



“信息化与信息社会”系列丛书编委会名单

编委会主任 曲维枝

编委会副主任 周宏仁 张尧学 徐 愈

编委会委员 何德全 邬贺铨 高新民 高世辑 张复良 刘希俭
刘小英 李国杰 秦 海 赵泽良 杜 链 朱森第
方欣欣 陈国青 李一军 李 琪 冯登国

编委会秘书处 廖 瑾 刘宪兰 刘 博 等

高等学校信息管理与信息系统专业系列教材编委会名单

专业编委会顾问 (以汉字拼音为序)

陈 静 杜 链 冯惠玲 高新民 黄梯云 刘希俭
王安耕 汪玉凯 王众托 邬贺铨 杨国勋 周汉华
周宏仁 朱森第

专业编委会主任 陈国青 李一军

专业编委会委员 (以汉字拼音为序)

陈国青 陈 禹 胡祥培 黄丽华 李 东 李一军
马费成 王刊良 杨善林

专业编委会秘书 闫相斌 卫 强

本书主审 王众托

普通高等教育“十二五”国家级规划教材

工业和信息化部“十二五”规划教材

“信息化与信息社会”系列丛书之
高等学校信息管理与信息系统专业系列教材

信 息 组 织

(第2版)

叶继元 主编

電子工業出版社·

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书系统讲述了信息组织的基本知识、经典理论,详细介绍了信息的描述、著录、标引和排序等方法,注重引入与信息组织有关的最新研究成果,如信息构建理论、本体论、搜索引擎、本体描述语言、SKOS 描述语言、学科信息门户等。全书分为 9 章,内容包括:信息组织概述、信息组织的理论与方法基础、信息描述语言、信息著录法、信息标引法、信息排检法、信息组织成果与工具、语义网环境下的信息组织、不同环境下的信息组织评价。每章正文之前有内容提要和本章重点,在每章后都附有本章小结和一定数量的讨论题,以满足教学和自学的需要。

本书具有定位明确、结构新颖合理、注重经典内容和社会实践等特点,可以作为高等院校信息管理与信息系统专业、图书馆学、情报学、档案学、编辑出版、博物馆学及相关信息专业的教材,也可以作为信息管理部门、图书情报界、各类与信息组织有关的机构或部门、专业工作者的参考书,还可以作为对信息组织感兴趣的读者的自学读物。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有,侵权必究。

图书在版编目(CIP)数据

信息组织 / 叶继元主编. —2 版. —北京: 电子工业出版社, 2015.3
高等学校信息管理与信息系统专业系列教材
ISBN 978-7-121-25097-2

I. ①信… II. ①叶… III. ①信息管理—高等学校—教材 IV. ①G203

中国版本图书馆 CIP 数据核字(2014)第 290638 号

策划编辑: 刘宪兰

责任编辑: 郝黎明

印 刷:

装 订:

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本: 787×1092 1/16 印张: 30 字数: 768 千字

版 次: 2010 年 2 月第 1 版

2015 年 3 月第 2 版

印 次: 2015 年 3 月第 1 次印刷

定 价: 58.00 元

凡所购买电子工业出版社图书有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系, 联系及邮购电话: (010) 88254888。

质量投诉请发邮件至 zltz@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线: (010) 88258888。



作者简介

叶继元，南京大学特聘教授，信息管理学院博士生导师；中国人文社会科学评价国家创新基地副主任；教育部首届、第二届社会科学委员会，国务院学位委员会第五届学科评议组成员；全国高校图书情报指导委员会期刊专业委员会副主任；中国索引学会副理事长；美国《历史文摘》(Historical Abstracts)编辑顾问(1991—1997)等；美国 Kansas 大学访问学者(1999—2000)；美国 Drexel 大学讲座教授、高级访问学者(2006)；国家社科基金重大项目首席专家。《核心期刊概论》、《学术规范通论》于 1998 年、2006 年分别获教育部人文社会科学优秀研究成果二等奖和江苏省哲学社会科学优秀研究成果二等奖；指导的博士论文《核心网站评选的理论与方法》获 2006 年全国优秀博士论文提名奖；组织全国外刊订购协调，取得重大经济和社会效益，1996 年获教育部有关部门通报表彰；《信息检索导论》获“十一五”国家级规划教材和高等学校信息管理示范教材。主要教学和研究领域包括文献信息资源建设与信息检索、期刊与文献计量学、学术评价与学术规范研究、图书馆学情报学实践与理论研究。



第2版 总序

信息化是世界经济和社会发展的必然趋势。近年来，在党中央、国务院的高度重视和正确领导下，我国信息化建设取得了积极进展，信息技术对提升工业技术水平、创新产业形态、推动经济社会发展发挥了重要作用。信息技术已成为经济增长的“倍增器”、发展方式的“转换器”、产业升级的“助推器”。

作为国家信息化领导小组的决策咨询机构，国家信息化专家咨询委员会按照党中央、国务院领导同志的要求，就我国信息化发展中的前瞻性、全局性和战略性的问题进行调查研究，提出政策建议和咨询意见。信息化所具有的知识密集的特点，决定了人力资本将成为国家在信息时代的核心竞争力。大量培养符合中国信息化发展需要的人才国家信息化发展的一个紧迫需求，也是我国推动经济发展方式转变，提高在信息时代参与国际竞争比较优势的关键。2006年5月，我国公布《2006—2010年国家信息化发展战略》，提出“提高国民信息技术应用能力，造就信息化人才队伍”是国家信息化推进的重点任务之一，并要求构建以学校教育为基础的信息化人才培养体系。

为了促进上述目标的实现，国家信息化专家咨询委员会致力于通过讲座、论坛、出版等各种方式推动信息化知识的宣传、教育和培训。2007年，国家信息化专家咨询委员会联合教育部、原国务院信息化工作办公室成立了“信息化与信息社会”系列丛书编委会，共同推动“信息化与信息社会”系列丛书的组织编写工作。编写该系列丛书的目的是，结合我国信息化发展的实际和需求，针对国家信息化人才教育和培养工作，有效梳理信息化的基本概念和知识体系，通过高校教师、信息化专家、学者与政府官员之间的相互交流和借鉴，充实我国信息化实践中的成功案例，进一步完善我国信息化教学的框架体系，提高我国信息化图书的理论和实践水平。毫无疑问，从国家信息化长远发展的角度来看，这是一项带有全局性、前瞻性和基础性的工作，是贯彻落实国家信息化发展战略的一个重要举措，对于推动国家的信息化人才教育和培养工作，加强我国信息化人才队伍的建设具有重要意义。

考虑到当时国家信息化人才培养的需求，各个专业和不同教育层次（博士生、硕士生、本科生）的需要，以及教材开发的难度和编写进度时间等问题，“信息化与信息社会”系列丛书编委会采取了集中全国优秀学者和教师，分期分批出版高质量的信息化教育丛书的方式，结合高校专业课程设置情况，在“十一五”期间，先后组织出版了“信息管理与信息系统”、“电子商务”、“信息安全”三套本科专业高等学校系列教材，受到高校相关专业学科及相关专业师生的热烈欢迎，并得到业内专家与教师的一致好评和高度评价。

但是，随着时间的推移和信息技术的快速发展，上述专业的教育面临着持续更新、不断完善的迫切要求，日新月异的技术发展及应用变迁也不断对新时期的建设和人才培养提出新要求。为此，“信息管理与信息系统”、“电子商务”、“信息安全”三个专业教育需以综合的视角和发展的眼光不断对自身进行调整和丰富，已出版的教材内容也需及时进行更新和调整，以满足需求。

这次，高等学校“信息管理与信息系统”、“电子商务”、“信息安全”三套系列教材的修

订是在涵盖第1版主题内容的基础上,进行的更新和调整。我们希望在内容构成上,既保持原第1版教材基础的经典内容,又要介绍主流的知识、方法和工具,以及最新的发展趋势,同时增加部分案例或实例,使每一本教材都有明确的定位,分别体现“信息管理与信息系统”、“电子商务”、“信息安全”三个专业领域的特征,并在结合我国信息化发展实际特点的同时,选择性地吸收国际上相关教材的成熟内容。

对于这次三套系列教材(以下简称系列教材)的修订,我们仍提出了基本要求,包括信息化的基本概念一定要准确、清晰,既要符合中国国情,又要与国际接轨;教材内容既要符合本科生课程设置的要求,又要紧跟技术发展的前沿,及时地把新技术、新趋势、新成果反映在教材中;教材还必须体现理论与实践的结合,要注意选取具有中国特色的成功案例和信息技术产品的应用实例,突出案例教学,力求生动活泼,达到帮助学生学以致用目的,等等。

为力争修订教材达到我们一贯秉承的精品要求,“信息化与信息社会”系列丛书编委会采用了多种手段和措施保证系列教材的质量。首先,在确定每本教材的第一作者的过程中引入了竞争机制,通过广泛征集、自我推荐和网上公示等形式,吸收优秀教师、企业人才和知名专家参与写作;其次,将国家信息化专家咨询委员会有关专家纳入各个专业编委会中,通过召开研讨会和广泛征求意见等多种方式,吸纳国家信息化一线专家、工作者的意见和建议;再次,要求各专业编委会对教材大纲、内容等进行严格的审核,并对每本教材配有一至两位审稿专家。

我们衷心期望,系列教材的修订能对我国信息化相应专业领域的教育发展和教学水平的提高有所裨益,对推动我国信息化的人才培养有所贡献。同时,我们也借系列教材修订出版的机会,向所有为系列教材的组织、构思、写作、审核、编辑、出版等做出贡献的专家学者、教师和工作人员表达我们最真诚的谢意!

应该看到,组织高校教师、专家学者、政府官员及出版部门共同合作,编写尚处于发展动态之中的新兴学科的高等学校教材,有待继续尝试和不断总结经验,也难免会出现这样或那样的缺点和问题。我们衷心希望使用该系列教材的教师和学生能够不吝赐教,帮助我们不断地提高系列教材的质量。

曲维枝

2013年11月1日



第1版 总序

信息化是世界经济和社会发展的必然趋势。近年来，在党中央、国务院的高度重视和正确领导下，我国信息化建设取得了积极进展。信息技术对提升工业技术水平、创新产业形态、推动经济社会发展发挥了重要作用。信息技术已成为经济增长的“倍增器”、发展方式的“转换器”、产业升级的“助推器”。

作为国家信息化领导小组的决策咨询机构，国家信息化专家咨询委员会一直在按照党中央、国务院领导同志的要求就信息化前瞻性、全局性和战略性的问题进行调查研究，提出政策建议和咨询意见。在做这些工作的过程中，我们愈发认识到，信息技术和信息化所具有的知识密集的特点，决定了人力资本将成为国家在信息时代的核心竞争力，大量培养符合中国信息化发展需要的人才已成为国家信息化发展的一个紧迫需求，成为我国应对当前严峻经济形势，推动经济发展方式转变，提高在信息时代参与国际竞争比较优势的关键。2006年5月，我国公布《2006—2010年国家信息化发展战略》，提出“提高国民信息技术应用能力，造就信息化人才队伍”是国家信息化推进的重点任务之一，并要求构建以学校教育为基础的信息化人才培养体系。

为了促进上述目标的实现，国家信息化专家咨询委员会一直致力于通过讲座、论坛、出版等各种方式推动信息化知识的宣传、教育和培训工作。2007年，国家信息化专家咨询委员会联合教育部、原国务院信息化工作办公室成立了“信息化与信息社会”系列丛书编委会，共同推动“信息化与信息社会”系列丛书的组织编写工作。编写该系列丛书的目的，是力图结合我国信息化发展的实际和需求，针对国家信息化人才教育和培养工作，有效梳理信息化的基本概念和知识体系，通过高校教师、信息化专家、学者与政府官员之间的相互交流和借鉴，充实我国信息化实践中的成功案例，进一步完善我国信息化教学的框架体系，提高我国信息化图书的理论和实践水平。毫无疑问，从国家信息化长远发展的角度来看，这是一项带有全局性、前瞻性和基础性的工作，是贯彻落实国家信息化发展战略的一个重要举措，对于推动国家的信息化人才教育和培养工作，加强我国信息化人才队伍的建设具有重要意义。

考虑到当前国家信息化人才培养的需求、各个专业 and 不同教育层次（博士生、硕士生、本科生）的需要，以及教材开发的难度和编写进度时间等问题，“信息化与信息社会”系列丛书编委会采取了集中全国优秀学者和教师、分期分批出版高质量的信息化教育丛书的方式，根据当前高校专业课程设置情况，先开发“信息管理与信息系统”、“电子商务”、“信息安全”三个本科专业高等学校系列教材，随后再根据我国信息化和高等学校相关专业发展的情况陆续开发其他专业和类别的图书。

对于新编的三套系列教材（以下简称系列教材），我们寄予了很大希望，也提出了基本要求，包括信息化的基本概念一定要准确、清晰，既要符合中国国情，又要与国际接轨；教材内容既要符合本科生课程设置的要求，又要紧跟技术发展的前沿，及时地把新技术、新趋势、新成果反映在教材中；教材还必须体现理论与实践的结合，要注意选取具有中国特色的成功案例和信息技术产品的应用实例，突出案例教学，力求生动活泼，达到帮助学生学以致

用的目的，等等。

为力争出版一批精品教材，“信息化与信息社会”系列丛书编委会采用了多种手段和措施保证系列教材的质量。首先，在确定每本教材的第一作者的过程中引入了竞争机制，通过广泛征集、自我推荐和网上公示等形式，吸收优秀教师、企业人才和知名专家参与写作；其次，将国家信息化专家咨询委员会有关专家纳入各个专业编委会中，通过召开研讨会和广泛征求意见等多种方式，吸纳国家信息化一线专家、工作者的意见和建议；最后，要求各专业编委会对教材大纲、内容等进行严格的审核，并对每一本教材配有一至两位审稿专家。

如今，我们很高兴地看到，在教育部和原国务院信息化工作办公室的支持下，通过许多高校教师、专家学者及电子工业出版社老师们的辛勤努力和付出，“信息化与信息社会”系列丛书中的三套系列教材即将陆续和读者见面。

我们衷心期望，系列教材的出版和使用能对我国信息化相应专业领域的教育发展和教学水平提高有所裨益，对推动我国信息化的人才培养有所贡献。同时，我们也借系列教材开始陆续出版的机会，向所有为系列教材的组织、构思、写作、审核、编辑、出版等做出贡献的专家学者、教师和工作人员表达我们最真诚的谢意！

应该看到，组织高校教师、专家学者、政府官员及出版部门共同合作，编写尚处于发展动态之中的新兴学科的高等学校教材，还是一个初步的尝试。其中，固然有许多的经验可以总结，也难免会出现这样那样的缺点和问题。我们衷心地希望使用系列教材的教师和学生能够不吝赐教，帮助我们不断地提高系列教材的质量。

曲维枝

2008年12月15日



第2版 序言

在移动计算、物联网、云计算等一系列新兴技术的支撑下,网络生活、社交媒体、协同创造、虚拟服务等新型应用模式持续拓展着人类创造和利用信息的范围与形式。这些日新月异的新兴技术与应用模式的涌现,使得全球数据量呈现前所未有的爆发式增长态势。同时,数据复杂性也急剧增加,其多样性(多源、异构、多模态和富媒体等)、低价值密度(信息不相关性和高“提纯”难度等)、实时性(流信息和连续商务等)特征日益显著。可以说我们已经进入“大数据”时代。数据已经渗透到每一个行业和领域,成为国家宏观调控和治理,社会各行各业管理和技术应用的基础和要素。

大数据时代的管理喻意可以从两个方面来概括,即“三个融合”和“三新”。“三个融合”指IT融合(信息技术与社会生活及企业业务的密不可分性)、内外融合(企业外部数据与内部数据整合的重要性)和价值融合(企业“造”与“用”价值创造的模式创新性)。这三个融合意味着:①越来越多的传统管理和决策成为基于数据分析的管理和决策(如数字化生存、数据运营、深度业务分析(Business Analytics, BA)核心能力等);②用户/公众创造内容(UGC/PGC)(如评论、口碑、商誉、舆情和社会网络等)成为企业活动的重要关注点;③企业的价值创造过程日益体现出“无形围绕有形”的互动(如“服务围绕产品”的业务拓展方式等)。而“三新”则指大数据时代催生的新模式、新业态和新人群。这意味着:①现有企业需要升级转型(如数据驱动的精益管理和模式创新等);②新业态在诞生和发展(如赛博空间生活和众包等);③信息社会中“移民”和“原住民”的多样化生存(如新型客户关系、新式企业文化和新颖行为特点等)。大数据时代管理喻意的上述两个方面反映了大数据时代管理理论和实践的变化特征,其中前者主要体现管理领域和视角上的变化,后者则主要体现管理主体和方式上的变化。

在我国信息化与工业化、城镇化和农业现代化同步发展的背景下,展望我国信息化发展的未来,信息技术应用将持续呈现出在物联网和智慧城市建设、云平台和大数据分析、新兴电子商务应用、企业信息化新拓展、绿色信息化路径等领域的主流现象和发展趋势,也为高等学校“信息管理与信息系统”专业建设和人才培养在新形势下带来新的挑战和机遇。

“信息管理与信息系统”作为一个快速更迭、动态演进的学科专业,必须以综合的视角和发展的眼光不断对自身进行调整和丰富,以适应新时代前进的步伐。高等学校信息管理与信息系统专业系列教材的第2版修订,就是希望通过更为系统化的逻辑体系和更具前瞻性的内容组织,帮助信息管理与信息系统专业相关领域的学生及实践者更好地理解现代信息系统在“造”(技术)和“用”(管理)维度上的分野和统一,掌握相关的基础知识和基本技能(特别包括企业进行数据运营、利用深度业务分析(BA)构建核心竞争能力方面的基础知识和技能)。

本次对高等学校信息管理与信息系统专业系列教材的修订,在基本保留第1版主要内容的框架基础上,仍然强调把握领域知识的“基础、主流与发展”的关系,并体现“管理与技术并重”的领域特征。同时,在整个系列和相关教材内容中,从领域发展与知识点的角度,

以不同程度和形式反映新技术时代的特点（如云计算和大数据这一新型计算模式）、IT 应用特征（如移动性、虚拟性、个性化、社会性和极端数据）、信息化拓展（如两化深度融合和企业外部数据分析）、新兴电子商务应用（如移动商务、社会化商务和 O2O）、搜索方法与服务（如关键词搜索与营销、信息检索与匹配）、IT 战略与管理（如服务管理、伙伴管理、业务安全管理和连续商务管理）等。我们希望通过系列教材专业编委会的共同努力，第 2 版系列教材能够成为高等学校信息管理与信息系统专业及相关专业学生循序渐进了解和掌握专业知识的系统性学习材料，成为大数据环境下从业人员及管理者的有益参考资料。

本系列教材的编写和修订得到了多方面的帮助与支持。在此，我们感谢国家信息化专家咨询委员会及高等学校信息管理与信息系统系列教材编委会专家们对教材体系设计的指导和建议，感谢教材编写者在时间和精力上的大量投入及所在单位给予的大力支持，感谢参与本系列教材研讨和编审的各位专家、学者的真知灼见！同时，我们对电子工业出版社在本系列教材整个出版过程中所做的努力深表谢意！

由于时间和水平有限，第 2 版系列教材在内容上肯定存在不足和不尽如人意之处，恳请广大读者批评指正。

高等学校信息管理与信息系统
专业系列教材编委会
2013 年 12 月于北京



第1版 序言

日新月异的技术发展及应用变迁不断给信息系统的建设者与管理者带来新的机遇和挑战。例如,以 Web 2.0 为代表的社会性网络应用的发展深层次地改变了人们的社会交往行为及协作式知识创造的形式,进而被引入企业经营活动中,创造出内部 Wiki (Internal Wiki)、预测市场 (Prediction Market) 等被称为 “Enterprise 2.0” 的新型应用,为企业知识管理和决策分析提供了更为丰富而强大的手段;以 “云计算” (Cloud Computing) 为代表的软件和服务平台服务技术,将 IT 外包潮流推向了一个新的阶段,像电力资源一样便捷易用的 IT 基础设施和计算能力已成为可能;以数据挖掘为代表的商务智能技术,使得信息资源的开发与利用在战略决策、运作管理、精准营销、个性化服务等各个领域发挥出难以想象的巨大威力。对于不断推陈出新的信息技术与信息系统应用的把握和驾驭能力,已成为现代企业及其他社会组织生存发展的关键要素。

2008 年中国互联网络信息中心 (CNNIC) 发布的《第 23 次中国互联网络发展状况统计报告》显示,我国的互联网用户数量已超过 2.98 亿人,互联网普及率达到 22.6%,网民规模全球第一。与 2000 年相比,我国互联网用户的数量增长了 12 倍。换句话说,在过去的 8 年间,有 2.7 亿中国人开始使用互联网。可以说,这样的增长速度是世界上任何其他国家所无法比拟的,并且可以预期,在今后的数年中,这种令人瞩目的增长速度仍将持续,甚至进一步加快。伴随着改革开放的不断深入,互联网的快速渗透推动着中国经济、社会环境大步迈向信息时代。从而,我国 “信息化” 进程的重心,也从企业生产活动的自动化,转向了全球化、个性化、虚拟化、智能化、社会化环境下的业务创新与管理提升。

长期以来,信息化建设一直是我国国家战略的重要组成部分,也是国家创新体系的重要平台。近年来,国家在中长期发展规划及一系列与发展战略相关的文件中充分强调了信息化、网络文化和电子商务的重要性,指出信息化是当今世界发展的大趋势,是推动经济社会发展和变革的重要力量。《2006—2020 年国家信息化发展战略》提出要能 “适应转变经济增长方式、全面建设小康社会的需要,更新发展理念,破解发展难题,创新发展模式”,这充分体现出信息化在我国经济、社会转型过程中的深远影响,同时也是对新时期信息化建设和人才培养的新要求。

在这样的形势下,信息管理与信息系统领域的专业人才,只有依靠开阔的视野和前瞻性的思维,才有可能在这迅猛的发展历程中紧跟时代的脚步,并抓住机遇做出开拓性的贡献。另一方面,信息时代的经营、管理人才及知识经济环境下各行各业的专业人才,也需要拥有对信息技术发展及其影响力的全面认识和充分的领悟,才能在各自的领域之中把握先机。

因此,信息管理与信息系统的专业教育也面临着持续更新、不断完善的迫切要求。我国信息系统相关专业的教育已经历了较长时间的发展,形成了较为完善的体系,其成效也已初步显现,为我国信息化建设培养了一大批骨干人才。但我们仍然应该清醒地意识到,作为一个快速更迭、动态演进的学科,信息管理与信息系统专业教育必须以综合的视角和发展的眼光不断对自身进行调整和丰富。本系列教材的编撰,就是希望能够通过更为系统化的逻辑体

系和更具前瞻性的内容组织，帮助信息管理与信息系统相关领域的学生及实践者更好地掌握现代信息系统建设与应用的基础知识和基本技能，同时了解技术发展的前沿和行业的最新动态，形成对新现象、新机遇、新挑战的敏锐洞察力。

本系列教材旨在体系设计上较全面地覆盖新时期信息管理与信息系统专业教育的各个知识层面，既包括宏观视角上对信息化相关知识的综合介绍，也包括对信息技术及信息系统应用发展前沿的深入剖析，同时还提供了对信息管理与信息系统建设各项核心任务的系统讲解。此外，对一些重要的信息系统应用形式也进行了重点讨论。本系列教材主题涵盖信息化概论、信息与知识管理、信息资源开发与管理、管理信息系统、商务智能原理与方法、决策支持系统、信息系统分析与设计、信息组织与检索、电子政务、电子商务、管理系统模拟、信息系统项目管理、信息系统运行与维护、信息系统安全等内容。在编写中注意把握领域知识上的“基础、主流与发展”的关系，体现“管理与技术并重”的领域特征。我们希望，这套系列教材能够成为相关专业学生循序渐进了解和掌握信息管理与信息系统专业知识的系统性学习材料，同时也成为知识经济环境下从业人员及管理者的有益参考资料。

作为普通高等教育“十一五”国家级规划教材，本系列教材的编写得到了多方面的帮助和支持。在此，我们感谢国家信息化专家咨询委员会及高等学校信息管理与信息系统系列教材编委会专家们对教材体系设计的指导和建议；感谢教材编写者及所在各单位的大力支持；感谢参与本系列教材研讨和编审的各位专家、学者的真知灼见。同时，我们对电子工业出版社在本系列教材编辑和出版过程中所做的各项工作深表谢意。

由于时间和水平有限，本系列教材难免存在不足之处，恳请广大读者批评指正。

高等学校信息管理与信息系统
专业系列教材编委会
2009年1月



第2版 前言

信息组织是关于信息资源或信息体的描述或著录、标引和排序的学问,其各种成果如书目、目录、索引、文摘、信息导航、搜索引擎等则是这门学问研究的主要对象,信息的序化及序化的程度是该领域要解决的主要问题。信息组织是一门涉及多种学科,如逻辑学、语言学、认知心理学、计算机科学、图书情报学等知识领域,同时又是一门实践性很强的课程。因此,信息组织是信息管理的基础,是信息管理领域核心课程之一,是所有从事信息管理的专业人员,不论是图书情报、编辑出版、博物档案人员,还是信息系统设计、电子商务、电子政务、数据库研制者,都必须掌握的基本知识。随着全球信息化、计算机化、网络化、自动化和智能化的迅速发展,信息组织的理论和方法也面临着新的挑战和机遇,如何坚守已有的信息组织原则,在此基础上拓展出新的未来的组织原则是目前国内外信息组织和信息管理界积极探讨的热点问题。

信息组织与文献组织、知识组织既有联系又有区别。其相同点在于三者都是对信息的外部特征和内容特征进行揭示、描述和整序,不同点在于文献组织尽管也有分类、主题对内容特征的揭示和排序,但侧重于信息的外部特征,主要以文献为单元来组织;信息组织侧重于信息的内容特征,以信息体中的内容为组织对象;知识组织则更加专注于信息的内容特征。可以将这三者看成既有联系又有区别的三个发展阶段。文献组织是信息组织的先驱,而知识组织则是信息组织发展的高级阶段。信息组织中的“信息”,不仅指文本信息,而且包括网络信息及其他数字信息,甚至也可理解成知识中的显性知识。从这个意义上说,信息组织起着承上启下、继往开来的作用,因为它既包括文献组织的固有内容,又涉及显性知识组织的新兴内容,也为探讨隐性知识的组织预留了空间。

近年来,国内外已正式出版的有关信息组织的教材共有十多部,这些教材各有其历史渊源和特点,本教材与同类教材相比,力求达到以下要求。

1. 定位明确

本书是为信息管理与信息系统专业本科生的培养需要而写的。众所周知,信息管理与信息系统专业是由五个专业合并而成的,因此本书所阐述的知识点力求符合五个专业的共性需求。

2. 结构合理

本书的内容结构安排力求具有新意。内容结构以信息组织的流程为主干,从一般到具体再到一般,强调章节之间的逻辑性和体系性。例如,各章节的安排是先概述信息组织的定义、作用、历史等,后讲述信息组织的理论和方法基础,接着从信息描述语言、信息著录法、信息标引法、信息排检法、信息组织成果与工具、语义网环境下的信息组织分别列章讲述,最后是对不同环境下的信息组织评价。既有信息的描述、著录,亦有信息的标引排序;既包含传统知识,亦包含最新的成熟内容;既有国内的内容,亦有国外相关教材的成熟内容的介绍;既有手工组织的内容,亦有自动组织的内容。全书的结构合理、清晰。

3. 注重经典内容

信息组织的内容丰富，为了便于本科生理解，本书注重基础的经典内容、基本概念和原理、主流的知识、方法和工具及信息与知识组织发展趋势的讲述。例如，注重讲述信息组织的基本原理、信息著录的标准、国内外主要的分类法、主题法和本体描述语言、SKOS 描述语言、目录、索引、数据库等的介绍，强调信息组织的一些原则、原理不会随着信息体的改变、组织形式的变化而失效，相反，会随着其变化而不断丰富。

4. 注重社会实践

随着社会信息化、网络化、知识化等的发展，信息组织的社会实践工具和成果不断丰富，本书及时吸收实践成果，注重技术和管理融合，避免单纯的理论叙述。例如，本书大量结合信息组织实例，如 Marc、元数据、网络数据库及 Google、雅虎等的最新发展，评价其得失。

本书修订始于 2013 年 1 月，编写者来自北京大学信息管理系、武汉大学信息管理学院、南京大学信息管理学院、南京理工大学信息管理系、南京农业大学信息管理系、安徽大学信息管理学院，编写者大部分都曾经参与过《信息组织》第一版的撰写，并曾出版、发表过有关信息组织的论著或教材，或者讲授信息组织有关课程，具有丰富的教学经验和实践经验。本书的写作大纲、全书的统稿、各章节的修改由叶继元负责。第 1 章由刘磊、冯英华编写；第 2 章由叶继元、刘丹、王晓艳编写；第 3 章第 1 节由薛春香编写，第二节由谭华军编写；第 3 节由岳泉编写；第 4 章第 1 节由谭华军编写，第 2 节、第 3 节由罗琳编写；第 5 章第 1 节由谭华军编写，第 2 节由岳泉编写，第 3 节由徐美凤编写；第 6 章由郭春侠、储节旺编写；第 7 章第 1 节、第 2 节由华薇娜编写；第 3 节、第 4 节、第 5 节由薛春香编写；第 8 章由欧石燕、王军编写；第 9 章由黄如花编写。

在本书编写过程中，陈国青教授、李一军教授、国家信息化专家咨询委员会周宏仁常务副主任、编委会秘书长杨春艳女士、电子工业出版社编辑刘宪兰女士给予了大力支持和帮助，刘丹博士在校对、编务方面做了大量工作，王雅戈为本书编制了书后索引，在此表示衷心的感谢。对本书引用论著、教材的作者、编者也一并表示衷心的感谢。

由于时间紧、编写人员较多、内容涉及面广，书中难免有不当之处，恳请读者指正。

叶继元
2014 年 5 月



第1版 前言

信息组织是关于信息资源或信息体的描述或著录、标引和排序的学问,其各种成果如书目、目录、索引、文摘、信息导航、搜索引擎等则是这门学问研究的主要对象,信息的序化及其序化的程度是该领域要解决的主要问题。信息组织是一门涉及多种学科如逻辑学、语言学、认知心理学、计算机科学、图书情报学等的知识领域,同时又是一门实践性很强的课程。因此,信息组织是信息管理的基础,是信息管理领域核心课程之一,是所有从事信息管理的专业人员,不论是图书情报、编辑出版、博物档案人员,还是信息系统设计、电子商务、电子政务、数据库研制者,都必须掌握的基本知识。随着全球信息化、计算机化、网络化、自动化和智能化的迅速发展,信息组织的理论和方法也面临着新的挑战和机遇,如何坚守已有的信息组织原则,在此基础上拓展出新的未来的组织原理、原则是目前国内外信息组织和信息管理界积极探讨的热点问题。

信息组织与文献组织、知识组织既有联系又有区别。其相同点在于三者都是对信息的外部特征和内容特征进行揭示、描述和整序,不同点在于文献组织尽管也有分类、主题对内容特征的揭示和排序,但侧重于信息的外部特征,主要以文献为单元来组织;信息组织侧重于信息的内容特征,以信息体中的内容为组织对象;知识组织则更加专注于信息的内容特征。可以将这三者看成既有联系又有区别的三个发展阶段。文献组织是信息组织的先驱,而知识组织则是信息组织发展的高级阶段。信息组织中的“信息”,不仅指文本信息,而且包括网络信息及其他数字信息,甚至也可理解成知识中的显性知识。从这个意义上说,信息组织起着承上启下、继往开来的作用,因为它既包括文献组织的固有内容,又涉及显性知识组织的新兴内容,也为探讨隐性知识的组织预留了空间。

近年来,国内外已正式出版的有关信息组织的教材共有十多部,这些教材各有其历史渊源和特点,本教材与同类教材相比,力求达到以下要求。

1. 定位明确

本书是为信息管理与信息系统专业本科生的培养需要而写的。众所周知,信息管理与信息系统专业是由五个专业合并而成的,因此本书所阐述的知识点力求符合五个专业的共性需求。

2. 结构合理

本书的内容结构安排力求具有新意。内容结构以信息组织的流程为主干,从一般到具体再到一般,强调章节之间的逻辑性和体系性。例如,各章节的安排是先概述信息组织的定义、作用、历史等,后讲述信息组织的理论和方法基础,接着从信息描述语言、信息著录法、信息整序法、信息排检法、信息组织成果的流程分别列章讲述,最后是对信息组织成果实例的评价。既有信息的描述、著录,亦有信息的标引排序;既包含传统知识,亦包含最新的成熟内容;既有国内的内容,亦有国外相关教材的成熟内容的介绍;既有手工组织的内容,亦有自动组织的内容。全书的结构合理、清晰。

3. 注重经典内容

信息组织的内容丰富，为了便于本科生理解，本书注重基础的经典内容、基本概念和原理、主流的知识、方法和工具及信息与知识组织发展趋势的讲述。例如，注重讲述信息组织的基本原理、信息著录的标准、国内外主要的分类法、主题法和本体描述语言、SKOS 描述语言、目录、索引、数据库等的介绍，强调信息组织的一些原则、原理不会随着信息体的改变、组织形式的变化而失效，相反，会随着其变化而不断丰富。

4. 注重社会实践

随着社会信息化、网络化、知识化等的发展，信息组织的社会实践工具和成果不断丰富，本书及时吸收实践成果，注重技术和管理融合，避免单纯的理论叙述。例如，本书大量结合信息组织实例，如 March、元数据、网络数据库及 Google、雅虎等的最新发展，评价其得失。

本书的编写始于 2007 年 9 月，编写者来自北京大学信息管理系、武汉大学信息管理学院、南京大学信息管理系、南京理工大学信息管理系、南京农业大学信息管理系、安徽大学信息管理系，编写者都曾出版、发表过有关信息组织的论著或教材，或者讲授信息组织有关课程，具有丰富的教学经验和实践经验。本书的写作大纲、全书的统稿、各章节的修改由叶继元负责。第 1 章由刘磊、冯英华编写；第 2 章由叶继元、刘丹、王晓艳编写；第 3 章第 1 节、第 2 节由薛春香编写，第 3 节由王军、王一丁编写；第 4 章第 1 节由谭华军编写，第 2 节、第 3 节由罗琳编写；第 5 章第 1 节由谭华军编写，第 2 节、第 3 节由岳泉编写，第 4 节由徐美凤编写；第 6 章由郭春侠、储节旺编写；第 7 章由华薇娜编写；第 8 章由黄如花编写。

在本书编写过程中，陈国青教授、李一军教授、国家信息化专家咨询委员会周宏仁常务副主任、编委会秘书长杨春艳女士、电子工业出版社编辑刘宪兰女士给予了大力支持和帮助，刘丹、袁曦临、刘宇、王晓艳、陈铭、李星星、赵青在校对、编务方面做了大量工作，王雅戈为本书编制了书后索引，在此表示衷心的感谢。对本书引用论著、教材的作者、编者也表示衷心的感谢。

由于时间紧、编写人员较多、内容涉及面广，书中难免有不当之处，恳请读者指正。

叶继元
2009 年 8 月

目 录

第 1 章 信息组织概述	1
1.1 信息组织的界定	2
1.1.1 信息组织的概念	2
1.1.2 信息组织的原理	4
1.1.3 信息组织的内容	6
1.2 信息组织的类型	7
1.2.1 按信息的表现形式划分	7
1.2.2 按信息的加工程度划分	8
1.2.3 按信息的传播载体划分	9
1.2.4 按信息的认识层次划分	11
1.2.5 按信息的存在环境划分	12
1.3 信息组织的作用	13
1.3.1 控制整序作用	13
1.3.2 提升品质作用	13
1.3.3 传播利用作用	14
1.3.4 节约成本作用	14
1.4 信息组织的发展	14
1.4.1 古代的信息组织	14
1.4.2 近代的信息组织	18
1.4.3 现代的信息组织	21
1.4.4 信息组织的未来发展趋势	29
本章小结	34
问题讨论	34
第 2 章 信息组织的理论与方法基础	35
2.1 信息组织的理论基础	36
2.1.1 有序化理论	36
2.1.2 信息构建理论	39
2.1.3 知识论	42
2.1.4 本体论	45
2.1.5 分形理论	47
2.2 信息组织的方法基础	47
2.2.1 语言学	47
2.2.2 逻辑学	49

2.2.3 知识分类学	53
本章小结	65
问题讨论	65
第 3 章 信息描述语言	67
3.1 信息描述语言概述	68
3.1.1 规范语言	68
3.1.2 自然语言	74
3.2 分类语言	81
3.2.1 分类法的原理	81
3.2.2 分类法的编制	85
3.2.3 国内外常用分类法介绍	101
3.3 主题语言	118
3.3.1 主题法概述	118
3.3.2 国内外常用主题词表介绍	125
本章小结	143
问题讨论	143
第 4 章 信息著录法	145
4.1 传统著录法	146
4.1.1 传统著录法概述	146
4.1.2 文献信息著录规则	149
4.2 机读目录著录法	161
4.2.1 MARC 在全球的发展概述	161
4.2.2 MARC 记录基本格式	163
4.2.3 MARC 著录的优缺点	175
4.3 元数据著录法	177
4.3.1 元数据简介	177
4.3.2 都柏林核心元数据集	179
4.3.3 元数据描述框架	186
4.3.4 其他元数据	192
本章小结	198
问题讨论	199
第 5 章 信息标引法	201
5.1 分类标引	202
5.1.1 分类标引要求	202
5.1.2 分类标引方法	202
5.1.3 分类标引规则	207
5.2 主题标引	212
5.2.1 主题标引概述	212
5.2.2 主题标引方法	214
5.2.3 主题标引规则	219

5.2.4	主题标引与分类标引的比较	221
5.2.5	关键词标引	222
5.3	自动分类与自动标引	225
5.3.1	自动分类概述	225
5.3.2	自动标引概述	230
5.3.3	西文信息自动标引技术	231
5.3.4	汉语信息自动标引技术	234
5.3.5	信息自动标引有待研究的问题	239
	本章小结	241
	问题讨论	242
第 6 章	信息排检法	243
6.1	字序法	244
6.1.1	汉字音序排检法	244
6.1.2	汉字形序排检法	248
6.1.3	外文字顺法	252
6.2	类序法	254
6.2.1	学科体系分类排检法	254
6.2.2	事物性质分类排检法	256
6.2.3	网络信息分类排检法	257
6.2.4	主题词排检法	258
6.2.5	网络信息关键词排检法	259
6.3	时序法	260
6.3.1	历法常识	261
6.3.2	中国古代的时序法	262
6.3.3	时序法的应用	266
6.4	地序法及其他排检法	266
6.4.1	地序法	266
6.4.2	其他排检法	267
6.4.3	计算机程序与动态信息排检法	268
	本章小结	269
	问题讨论	270
第 7 章	信息组织成果与工具	271
7.1	目录	272
7.1.1	目录的类型	272
7.1.2	目录举要	273
7.2	索引与文摘	279
7.2.1	索引的概念	279
7.2.2	索引的类型	279
7.2.3	索引举要	283
7.2.4	文摘	290
7.3	全文数据库	291

7.3.1	全文数据库概述	291
7.3.2	全文数据库开发	293
7.3.3	全文数据库举要	295
7.4	异构数据库整合与导航	299
7.4.1	异构数据库整合	299
7.4.2	数据库导航	304
7.4.3	学科信息门户	308
7.5	搜索引擎	311
7.5.1	搜索引擎概述	311
7.5.2	搜索引擎工作机制	314
7.5.3	搜索引擎举要	315
7.5.4	搜索引擎的发展方向	316
	本章小结	317
	问题讨论	318
第 8 章	语义网环境下的信息组织	319
8.1	语义网概述	320
8.2	语义网信息描述与表示	321
8.2.1	RDF 简介	321
8.2.2	RDF 序列化表示格式	322
8.2.3	RDF 评价	324
8.3	语义网信息组织模式	324
8.3.1	本体简介	324
8.3.2	本体的类型	325
8.3.3	本体的功能	325
8.3.4	本体与传统情报检索语言的比较	326
8.3.5	本体的构建	326
8.3.6	本体描述语言	338
8.3.7	基于本体的信息组织实例	345
8.4	基于语义网的网络知识组织系统	347
8.4.1	SKOS 语言简介	347
8.4.2	SKOS-XL 语言简介	352
8.4.3	SKOS 语言应用实例	353
8.4.4	术语注册与术语服务	360
8.5	语义网信息组织方法	364
8.5.1	关联数据简介	364
8.5.2	关联数据中资源的命名及访问机制	364
8.5.3	关联数据中资源命名原则	365
8.5.4	关联数据发布方法	366
8.5.5	关联数据应用实例	367
	本章小结	370
	问题讨论	371

第 9 章 不同环境下的信息组织评价	373
9.1 非网络环境下的信息组织评价	374
9.1.1 印刷型文献的信息组织评价	374
9.1.2 多媒体信息组织评价	383
9.2 网络环境下的信息组织评价	394
9.2.1 网站信息资源组织评价	394
9.2.2 搜索引擎信息资源组织评价	399
9.2.3 学科信息门户资源组织评价	406
9.2.4 Web2.0 环境下信息组织评价	412
9.2.5 数字图书馆信息资源组织的实例与评价	421
本章小结	428
问题讨论	429
内容索引	431
参考文献	449



第1章

信息组织概述


本章引言

人类自产生以来，就在信息的海洋中生活了。例如，原始人在森林中搜寻野果和野兽的信息，探悉各种猎物的信息；人类很早就懂得利用信息的一些性质来达到特定的目的。例如，结绳记事就利用了信息的可存储性。

在当代信息社会中，信息已成为人所共知的流行词，日益受到人们的重视，它同能源和物质一起被称做是人类社会生产与生活必不可少的三大资源。随着现代科学技术特别是计算机技术、网络技术的迅猛发展，社会已经逐渐从信息社会向知识社会迈进，信息资源的开发和利用日益受到人们的广泛关注。

目前，信息资源的特点是海量、类多、源广，特别是计算机网络的逐步普及、数字文本复制的便利和自由发表的实现，使得信息资源数量急剧增加，但是社会信息量的增长并不意味着用户获取的信息量的增长，恰恰相反，无序的信息资源不仅无助于信息资源的使用，反而会加剧信息增长与使用的矛盾。人们生动地称这种情况是“信息超载，知识缺乏”。越来越多的人认识到，“原始信息本身并不能产生价值”，只有将其有效组织，按特定的需要集中和揭示，才能产生价值。要有效开发利用信息资源，必须采用相应的方法加以控制和处理，信息组织是关键措施之一。

本章重点

- 信息组织的概念；
 - 信息组织的原理；
 - 信息组织的类型；
 - 信息组织的作用；
 - 信息组织的发展。
- 

1.1 信息组织的界定

从某种角度来说，信息是推动人类社会发展的直接动力，这种动力在人类认识和改变世界过程中起到至关重要的作用，已成为现代社会生产力的基本要素。为了更好地发挥信息的积极作用，高效的信息管理则势在必行。在技术层面，信息管理的主要目的是通过手工或机械、智能（特别是计算机）方式对信息进行收集、甄别、加工、处理、存储，使之序化，便于快速检索并提交至有特定需求的用户。而对信息的加工、处理等序化工作则是信息组织的核心内容。因此，信息组织是信息管理的有机构成和重要环节。

现代社会信息的两个重要特征是庞杂性和分散性，人们的信息需求呈现出领域的高度选择性、内容的优质性、格式的易用性和时间的紧迫性等特点。这两者之间的联系及矛盾的调解依赖于信息组织。

信息组织是一个信息重构和增值过程。在这个过程中，杂乱无章的原始信息变成一个有序的、优质的信息集成系统，一个相对“粗放”型的信息贫集将转化为一个“集约”型的信息富集，并为信息的进一步增值（如信息的分析研究）奠定基础。

1.1.1 信息组织的概念

信息组织是随着社会信息化而出现的一个国内外同行均使用的术语。人类的生存与发展离不开信息，早期的信息获取主要靠人与人之间的直接交往。有了文字记载以后，人类在社会生产和生活中则越来越多地依赖于文献信息。随着文献的大量出现和急剧增长，便有了文献的整理加工工作，被习惯称作文献组织或情报组织，它为人们便捷获取特定文献信息提供了可能。

随着现代科学技术的发展，人类社会逐渐从信息社会向知识社会迈进，信息资源的开发和利用日益受到重视。目前信息资源的特点是数量大、种类多、来源广、格式杂，特别是计算机网络的全球普及、数字信息复制的便利和自由发表的实现进一步显化这些特点。但是，社会信息量的增长与用户获取和使用的信息量的增长并不具备同步性，恰恰相反，无序的信息资源不仅无助于信息资源的使用，反而会为用户获取相关信息设置障碍，加剧信息增长与使用的矛盾。人们生动地称这种现象是“信息超载，知识缺乏”。越来越多的人认识到“原始信息本身并不能产生价值”，只有将其有效组织和整合，按特定的需要集中和揭示，才能产生价值。从用户需求角度来看，多角度、多层次的信息组织将有利于满足用户多元化的信息需求。因此，要有效开发和利用信息资源，必须采用相应的方法对社会信息加以控制 and 处理，信息组织是其中的关键措施之一。

什么是信息组织？

要正确了解信息组织的概念，首先必须了解数据、信息、信息体和知识的含义及它们之间的区别。所谓数据，一般是指经过直接观察获得的事实，它是无生命的、未经过处理的。确切地说，它是“以正式的、适合人类或自动方式交流、转换、处理的对于事实、概念、程序的表述”；信息则是经过人类处理的数据，是人们通过惯常方式赋予数据意义得到的结果，是接收者对数据背景和规则的解读。需要明确的是，本书讨论的信息组织对象并不涵盖所有的信息，而是指可以组织的记录形式的信息、数据和知识，即信息体。在大数据时代，信息体包括的不仅仅是文本的数据、信息和知识，也包括非文本的数据、信息和知识（如音频、视频、图像、图形和网页）。尽管在现实生活中，数据和信息两个词经常替换使用。但严格地说，数据和信息是两个不同的概念。数据是按照一定规则排列组合记录信息的物理符号，

是信息和知识的基础。信息是数据载荷的内容,同一信息的数据表现形式可以多种多样。信息与经验(知识准备)相结合则可转化为知识。知识是信息接收者通过对信息的提炼和推理而获得的正确结论。在一定的语境下,信息和知识也可以互换。从这个意义上说,我们日常所说的许多信息系统和知识系统在本质上都是数据系统。

另外,信息组织与知识组织是既有联系又有区别的两个概念。信息组织是由文献组织发展而来的关于信息的组织与检索系统。而文献是指以硬载体(如纸张)为依托的信息。但由于现代信息数量急剧膨胀,信息类型日趋复杂化(如脱离纸张的计算机数字信息),传统的文献组织工作已经无法对信息加以有效地控制和管理。随着信息技术的飞速发展、计算机和网络的广泛普及,一方面人们开始将原始文献信息集合转化为有序、优质的计算机信息系统的数据库或文献信息资源的索引;另一方面,人们又面临着如何对直接产生于计算机或来自网络的数字信息进行组织的问题。因此,以印刷型信息为对象的文献组织必然发展成为以各种类型信息为对象的信息组织。知识组织是在文献分类法的基础上发展起来的,是关于知识的组织与检索系统,是现代网络信息环境下获取与利用知识的所有手段、技术与能力的总和,与文献组织和信息组织关系密切。信息组织对信息的有序排列是知识组织的原料基础,知识组织是信息组织发展到一定阶段的必然产物,比信息组织的内涵更深刻、丰富和明确,更能反映用户需求的实质,但其任务也更艰巨。

目前国内学者对于信息组织定义的表述各不相同,比较有代表性的有以下几种:马张华认为“信息组织,亦称为信息资源组织,是根据信息检索的需要,以文本及各种类型的信息源为对象,通过对其内容特征等的分析、选择、标引、处理,使其成为有序化集合的活动。”马费成认为:“信息组织也称信息整序,是利用一定的规则、方法和技术对信息的外部特征和内容特征进行揭示和描述,并按给定的参数和序列公式排列,使信息从无序集合转换为有序集合的过程。信息组织通过人工和机器干预,使信息有序增值,进而提供有效利用。”柯平等认为“信息组织是根据信息的内容特征和外部特征,采用一定的原则和方法,对信息进行加工处理,使之有序的过程。”周宁认为“信息组织是对信息资源对象进行收集、加工、整合、存储,使之有序化、系统化的过程。”戴维民认为:“信息组织是为了方便人们检索、获取信息而将庞杂、无序的信息进行系统化和有序化的过程。从广义上讲,信息组织的内容包括信息搜集与选择、信息分析与揭示、信息描述与加工,信息整理与存储。”储节旺认为“信息组织就是人们根据信息本身特点,运用适宜的工具和方法,依据一定的标准或规则,对其进行加工整理,排序组合,使之有序化、系统化、规律化、高级化,增强信息对象的表现效能和运用效能,以满足人们信息需求的过程和获得。”

以上定义虽然表述不太一样,但实质上是一致的。在社会网络环境下,考虑到用户已经参与到信息组织中来。我们认为,信息组织是以用户需求为导向,依据信息体自身的属性特征,信息工作者或用户按照一定的原则、方法和技术,将杂乱无章的信息整理成为有序的信息集合的活动和过程。信息组织的结果是形成各种方便用户利用的有序化的信息检索系统,从而达到信息增值的目的。信息组织是信息管理活动的核心和基本环节。

社会信息作为一种公共交流的知识,是通过不同的形式进行传播的,包括口头方式、纸质方式、数字方式等。各种信息载体都是信息存在的形式,同时也是交流的主要形式。以记录形式进行信息交流,能够克服面对面直接交流的时空局限,使得人类社会可以在继承已有经验的基础上向前发展,对人类社会的发展意义重大,是人类社会进步的重要条件。各种类型的信息资源既是产生公共知识的基础,又是社会发展的智力资源,是信息组织的主要处理对象。

需要说明的是,作为信息组织的处理对象,信息、信息资源或信息体在本教材中与文献

或资料的含义接近，通常是指一切以记录形式存在的信息载体，比较而言，只是更侧重于对新型媒体形式的强调。信息范围广泛、种类繁多，就信息的基本类型而言，它可以是图书、期刊、报纸、会议文献、标准文献和档案等各种类型；就存在的形式而言，它可以是各种传统的印刷型信息资源，也可以是缩微型、机读型资源及实体/非实体信息资源——如网络信息资源（或称数字信息资源）等。信息组织就是以各种信息媒体形式为对象进行的组织。就其处理的单元而言，信息组织存在着多种不同的层次：它可以直接以信息资源的存在单元为处理对象。例如，可以以某一图书、期刊、网站等为处理单元；也可以以期刊、报刊、网站中的个体，如论文、新闻、网站中的构成部分等为处理单元；还可以直接以其中的信息成分为处理单元，进行组织和揭示。本书主要讲述记录型信息，包括网络信息组织（数字信息组织）。

需要强调的是，信息组织是信息检索与利用的基础。没有它，就不会有信息检索系统，也不会有信息检索，从而也就谈不上对信息的有效利用。信息组织为人们获取信息提供了方法和工具。信息组织的目的是建立起信息资源收藏系统和检索工具，方便信息资源的开发和利用。信息组织是一种为了信息检索的需要对信息资源进行有序化组织的活动，它是与信息检索活动密切联系的。

信息组织是信息管理活动的必然要求，其起源在于信息本身的自然无序状态。序是事物的一种结构形态，是指事物或系统的各个结构要素之间的相互关系及这种关系在时间和空间中的表现，即事物发展中的时间序列及排列组合、聚类状态、结构层次等空间序列。当事物结构要素具有某种约束性且在时间序列和空间序列呈现某种规律性时，这一事物就处于有序状态；反之，则处于无序状态。

信息的无序状态有两种类型：一是信息内容无序，即组成信息的各个语言要素，如字、词、句、段落或章节等处于一种杂乱无章的状态，或是组成信息的各个内容要素，如命题、观点、认识、推理等处于一种分散无序的状态，或是组成信息的各个载体要素处于一种零散错位的状态，这种无序状态下的信息无法准确、科学地表述信息内涵、形成信息实体，也就无法被信息用户理解和应用；二是信息体及其相互之间的组织，即不同信息个体处于彼此毫无关联的自然状态，缺乏科学稳定的框架体系，没有信息的深层次加工，无法融合成一个可以科学地把握其信息内容、有效地查检其信息内容、完善地维护其信息形态、良好地排列其信息形态、充分地实现其信息价值的有序的信息集合体。

显然，本书中的信息组织所要解决的主要是第二种信息无序问题，第一种信息无序问题主要通过信息生产和信息搜集工作来解决。也就是说，信息组织是在信息搜集基础之上进行的信息系统的信息整理和序化工作。一般地，信息处于有序状态或无序状态是就一定参考系而言的。

信息组织是由来已久的一种人类社会实践活动。在其发展过程中，不断从相关学科的理论和方法中汲取营养，使自身逐渐得到完善。

1.1.2 信息组织的原理

在信息组织过程中，信息整序意义重大。信息整序既是理论问题，又是方法问题。整序是信息组织工作的基本手段和直接目的。众多的、杂乱分布的信息只有经过整序，才能得以方便地、有效地利用，才能最大程度发挥信息效能。

信息组织的基本原理是：如果有若干自然状态的无序信息资料（如若干图书或学位论文），若将这些信息资料按照其某种属性特征（如著者、题名、类属、主题）排列成一个序列，并且需要使用信息资料的用户能将自己的信息需求转换成相应的信息资料属性特征，并在排列后的信息资料序列中找到自己所需要的信息资料，则称这些信息资料是有序的。按照

信息属性特征排列信息的工作称为信息整序。

信息整序通常是针对信息记录的载体而进行的,然而,为了信息组织和检索的方便,信息组织的直接对象往往是信息载体的替代记录(亦称元数据)。替代记录反映了原信息载体的主要外部属性特征和内部属性特征,但在形式上却比原始信息载体要简洁得多。如图 1.1 中《国家数字图书馆服务框架研究》一书的在版编目数据就是该书的替代记录和元数据;又如图 1.2 中利用百度搜索该书的网页快照也是该书的替代记录。

国家数字图书馆服务框架研究/张炜主编.—北京:国家图书馆出版社,2012.5
ISBN 978-7-5013-4708

I. ①国… II. ①张… III. ①中国国家图书馆—数字图书馆—图书馆服务—研究 IV. ①G259.251

图 1.1 替代记录之一:图书在版编目数据

信息组织是由信息著录标引(信息描述揭示)和信息序化两个工作环节组成的。信息著录实际上是对原始信息的外部属性特征(题名、著者、出处等)和内部属性特征(类属、主题、摘要等)进行描述的过程;信息标引是给出信息内容标识(如分类号、主题词等)的揭示过程。著录标引的结果是将原始信息制成它的替代记录——二次信息(元数据)。信息序化则是将所有替代信息按照其某种外部特征(如著者、题名)和内容标识(如分类号、主题词)进行有规律的组织排列,从而构成某种序列(亦称某种目录或索引,如著者目录或索引),各种序列(目录索引)制作完成并存储以后,就形成了比较完整的检索系统。



图 1.2 替代记录之二:百度搜索网页快照

替代记录和原始信息之间存在着密切的关系。替代记录在信息组织和检索中的作用可通过图 1.3 来说明。信息组织与检索的实际过程是:信息工作者(含部分用户)在拿到原始信

息载体后，首先根据描述著录标引工具对其加工，形成替代记录，并将替代记录存储在检索系统中。检索时用户（或检索操作员）首先根据著录标引工具对用户的信息提问进行加工，形成替代问题。检索的核心过程是通过替代问题和替代记录的匹配比较进行的。在找到与替代问题相关的替代记录之后，再根据替代记录与原始信息的对应关系，查找到原始信息，以满足用户的原始问题。在网络信息组织中，这种替代机制同样存在。例如，在图 1.4 中，我们可以通过搜索引擎概要结构图，了解搜索引擎内部的信息组织和替代机制。

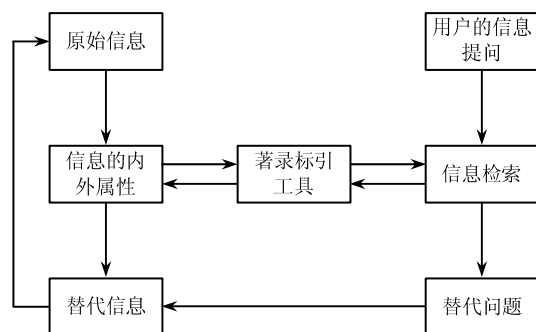


图 1.3 替代记录及其作用

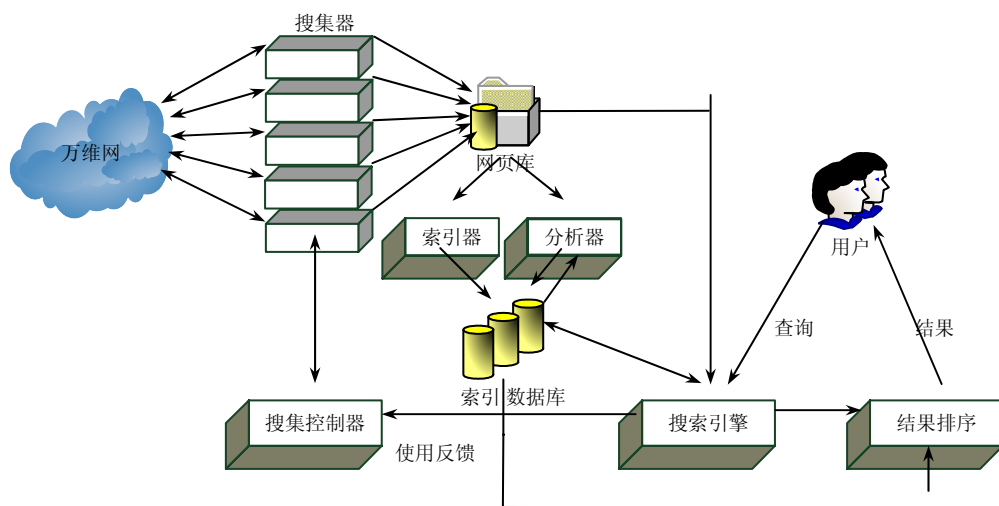


图 1.4 搜索引擎概要结构图

1.1.3 信息组织的内容

一般而言，信息组织包括信息筛选、信息分析、信息描述与揭示、信息序化和存储等内容。其中，信息选择与信息分析是信息组织的基础准备工作，信息描述与揭示是信息组织的核心，信息序化和存储是前面几个环节形成的有序信息集合的空间组织。由于信息组织通过对信息资源的序化和质化，最终实现向用户提供优质信息资源服务的目的。因此广义上，信息组织也可以归入信息服务的范畴。事实上，在社会网络环境下，信息组织与信息服务已经逐渐融为一体。

1. 信息筛选

信息筛选是按照一定的判别标准从搜集到的、处于无序或部分无序状态的信息流中甄选出有用信息并剔除无用信息的信息活动，它是整个信息组织过程的第一步。

2. 信息分析

信息分析是在信息筛选的基础上，按照特定的逻辑关系从语义、语用和语法上对信息资源的内外部信息特征进行细化、挖掘、加工整理并归类的信息活动。它是信息描述与揭示的前提和基础，直接影响着信息组织的质量。

3. 信息描述与揭示

信息描述是根据一定的描述规则和技术标准，对信息资源的部分主题内容、形式特征及物质形态等描述维度进行分析、选择、记录的过程；信息揭示是对信息资源的内容特征进行深层揭示并转换成主题标识系统的过程。信息描述与揭示主要分为两种类型：一是著录，主要描述文献信息的形式特征及部分内容特征；二是标引，主要揭示文献信息的内容特征。信息描述与揭示是信息组织的重要内容，在信息组织中起着至关重要的作用。

4. 信息整序与存储

信息存储是将经过加工整序后的信息资源按照一定的格式与顺序存储在特定的载体中的信息活动。信息存储是为了信息管理者和信息消费者快速、准确地识别、定位和检索信息，各种文献信息检索工具、光盘检索系统、网络信息检索工具等都是信息存储的重要方式。

1.2 信息组织的类型

信息源是人们生产生活、科学研究和其他一切活动中所产生的成果和各种原始记录，以及对这些成果和原始记录加工整理得到的成品。信息源是信息组织的对象，信息组织的类型基于信息或信息源的分类。本文按照信息的表现形式、信息的加工程度、信息的传播载体、信息的认识层次和按信息的存在环境对信息组织的类型进行细分。

1.2.1 按信息的表现形式划分

1. 文字信息组织

文字是人们为了记录事物、实现信息交流与通信联系所创造和约定的一种形象符号系统。广义的文字还包括各种编码，如 ASCII 码、国际电报与单元代码及计算机中的二进制数字编码等都是一些符号的约定。这些文字、符号、代码均是信息的表述形式，其内容再现于它们的结构属性中。如基本笔画的不同组合、字和字母的不同组合、二进制码“0”和“1”的不同排列等，分别代表不同的信息内容。这些文字信息是通过文字所表达的信息内容的语义、语用和语法来组织的。

2. 图像信息组织

图像是对客观对象的一种相似性的、生动性的描述或写真，是人类社会活动中最常用的信息载体，也是人们最主要的信息源。图像作为一种视角信息，它比文本信息更直观。图像本身具有分辨率、形状、大小、颜色、深度、饱和度、亮度及存储格式等属性特征。通常情况下，可通过表现图像的各种属性按照视觉的需要来组织图像信息，有时也按照对图像的存储格式的需要来组织。

例如，目前，为了提高网速，Web 网页上的图像一般都使用图形文件交换格式（GIF，Graphics Interchange Format）定义的.gif 格式文件和联合图像专家组（简称 JPEG）定义的.jpg 格式文件，有时也使用可移植的网络图像文件格式（简称 PNG）定义的.png 格式文件。使用.gif 格式存储的图像是无损压缩的图像，.png 格式是作为.gif 的无专利替代品开发的，在 World Wide Web 上无损压缩和显示图像；使用 JPG 格式的图像是有损压缩的图像，不过某些 Web 浏览器不支持.png 图像，而大多数 Web 浏览器都不需要调用其他的图像处理程序就能直接显示用.gif 和.jpg 这两种格式存储的图像。通常情况下，连续色调图像（如照片）应压缩为.jpeg 文件。具有单调颜色或锐化边缘及清晰细节的插图（如文字）应压缩为.gif 或.png 格式的。

3. 音频信息组织

声音和语言是人类交流时最普遍使用的方式，语音的基本参数包括基音周期、共振峰、语音谱、声强等，这些参数由声源、声道和放射机构（嘴唇和鼻孔）导出。声音可以进行编码和压缩，可以通过 A/D 转换将声音数字化并输入到计算机中处理，再通过 D/A 转换输出。我们一般按照听觉的需要或声音的记录方式来组织音频信息。

4. 视频信息组织

一般意义上，视频信息可分解为一系列静态图像信息和声音信息，可以看成是无数张图像帧按照一定的流向（顺序）排列并外加音频组配而成，对视频信息的组织实际上就是对图像和音频信息的组织。视频信息由于具有时空冗余性，压缩效果受到较大局限，最常用的视频信息压缩技术为 DVD、MPEG-2（卫星直播电视用）和 MPEG-4（互联网传输用）。从单个视频整合的角度来说，视频信息组织主要通过视频编辑软件，如 Premiere、iMovie 和 Cyberlink 等，可以实现对视频信息分线性地分割、合并、转换、输入、输出，集合图片、音乐、视频、字母等信息。而一般意义上对视频信息进行信息组织主要是从视频信息数据中抽取各种标识性辅助信息，如视频所有者、主题标签、时间长短、容量大小、画面清晰度等属性信息达到特定目的的视频信息有序化。目前基于内容的视频信息组织已被广泛应用。

1.2.2 按信息的加工程度划分

1. 一次信息组织

一次信息是人们研究或创造性活动成果的直接记录，一般指公开出版的图书、期刊论文、研究报告、会议文献、学位论文、专利文献和计算机网络上未经过再加工的数字化信息等。一次信息组织是采用一定的逻辑结构和语言规则，将大量零散的信息单元组织成能够反映信息所有者对事物和概念认识的一次信息（如图书、论文等）的信息组织方式。当前数字格式的一次信息大量存在，如计算机中的文本文件、图形文件、音频/视频文件、互联网的网页等。一次信息一般具有面广、量大、分散无序和良莠不齐等特点，有时很难获取，即使是在网上搜索，其查准率也非常低，逻辑相关性非常差。因此一次信息一般都需要进行信息整序才能方便用户使用。

2. 二次信息组织

二次信息是对一次信息加工后产生的信息，如目录、题录、简介、文摘、索引和书目数据库等。二次信息组织是在一次信息的基础上加工整序的信息组织方法，主要采用选择、提炼和浓缩等方法，编撰和建立各种形式的目录、文摘、索引等检索工具与存取系统，其主要功能是规范和控制信息的流向，组织序化文本与数字化信息。二次信息组织的目的是建立各类信息检索系统，它通过对一次信息的内容属性和外部特征进行有效揭示、描述、排序和存

储来实现。信息的外部特征主要指信息记录的载体直接反映出来的特征,如题名、著者、信息类型、语言、时间和地点等;信息的内部属性主要指信息记录所载知识内容的学科类属和主题属性。二次信息的形成是信息从分散、无序到集中、有序的控制过程。它的重要作用不仅在于报道,更重要的是为查找一次信息提供线索。

二次信息组织的产品很多,这些产品是人类分享信息的主要工具。它们可以以印刷型文献的形式出现,如传统的书目、索引、文摘等;也可以以数字化的方式出现,如各类书目型数据库(包括题录型和文摘型),互联网上的搜索目录和搜索引擎都是提供一次信息线索(包括网址和网页)的二次信息检索工具。

3. 三次信息组织

三次信息是根据特定的需求,对一次信息和二次信息进行加工、分析、改编、重组和综合概括生成的信息,根据其编制目的与方式,可将三次信息产品区分为综合(或专题)研究型 and 参考工具型两大类,前者如专题述评、总结报告、动态综述、信息预测等,后者如手册、百科全书、年鉴、指南等。三次信息组织是在一次、二次信息组织的基础上,根据某一学科、专业、主题或人物(类型)等需求,鉴别收集相关一次信息和二次信息进行,同时结合相关的数据统计与分析、数据挖掘方法和工具,以整序筛选浓缩与综合概况形成信息综述、信息汇编、信息述评或专题报告等的高层次信息组织方法。经过三次信息组织得到的信息产品在科技研发、企业竞争和政府决策等诸多领域中发挥着重大作用。

1.2.3 按信息的传播载体划分

1. 文本信息组织

一般来说,对文本的解释包括三个方面:在语言文学领域里,文本是一系列语句串联而成的连贯序列,通常是具有完整、系统含义的一个句子或多个句子的组合,一个文本可以是一个句子、一个段落或者是一个篇章;印刷型文献也是文本信息,包括图书、期刊、会议资料、专门报告、专利资料、政府出版物、学位论文、产品样本、档案、标准和报纸等;在计算机领域,文本主要指文本文档,是用于记载和储存文字信息,常见文本文档的扩展名有.txt、.doc、.docx和.wps等。显然,本书所涉及的文本信息组织的对象是后两者。

印刷型文本信息组织以分类组织法和主题组织法为主要形式。分类组织法是语法信息组织和语义信息组织的综合,从学科角度集约信息,便于族性检索。主题组织法是建立在自然语言基础之上的语义信息与语法信息组织的综合,其词族索引和范畴索引展现了主题词之间的等级关系和学科关系,属于语义信息组织,而附表和英汉对照索引则体现了语法信息的关系,主题组织法从事物角度集中信息,便于特性检索。计算机文本文档的信息组织则以计算机系统编码系统为准,主要通过对文本文档题名、作者、时间、格式、大小和存储地址等属性进行标识。

2. 多媒体信息组织

多媒体信息组织是指利用多媒体工具、多媒体语言工具等手段以及各种信息组织方法对文本、图形、声音、视频等多媒体信息进行集中处理,以增强信息的表现力,提高信息的使用价值的获得。多媒体信息组织广泛存在于电子出版物、多媒体教学软件等的设计和制作以及网络信息管理中。多媒体著作工具是能够集成处理和统一管理多媒体信息,使之能够根据用户的需要生成多媒体应用系统的工具软件,也有人称它为多媒体创作工具或多媒体写作工具等。多媒体著作工具提供了组织和编辑多媒体应用系统各媒体元素所需的框架,用多媒体著作工具设计交互性用户界面,将各种多媒体信息组合成一个连贯的节目。

3. 富媒体信息组织

富媒体（Rich Media）是指具有动画、声音、视频和/或交互性的信息传播方法，包含下列常见的形式之一或者几种的组合：流媒体、声音、Flash 以及 Java、Javascript 和 Dhtml 等程序设计语言。富媒体信息广泛应用于各种网络服务中，如网站设计、电子邮件、Banner、Button、弹出式广告和插播式广告等。富媒体信息组织基于网络信息资源和网络信息技术，按照特定原则和方法收集、选择、描述、标引和存储富媒体信息资源，以更好地序化和优化富媒体信息资源。富媒体信息组织方法主要包括以下几种。

（1）超文本法。超文本是一种新型的信息组织方法，超文本技术的一大特征是信息的非线性排列，它以节点为基本单位，节点间以链相连，将信息组织为某种网状结构。使用户可以从任一节点开始，根据网络中信息间的链接，从不同角度浏览和查询信息。超文本组织方法所提供的非顺序性的浏览功能，比传统的信息组织方式更加灵活、方便，且符合人们的联想思维方式。超文本技术的另一大特征是信息表达形式的多样性。超文本信息可以是文字、图形、图像、声音和动画等多种媒体形式，因此可称之为“超媒体”。

（2）自由文本法。该方法主要用于全文数据库的组织，是对非结构化的文本信息进行组织和处理的一种方式。它无须前控，不必用规范化语言对信息进行复杂的前处理。它不是对富媒体信息特征的格式化描述，而是用自然语言深入揭示其知识单元，根据富媒体的自然状况直接设置检索点。它能够完整地反映出富媒体的全貌，通过计算机自动进行文献处理和组织。基于全文数据库的全文检索可以将任意字符作为检索标识，用户用自然语言即可直接检索未经标引的原始信息。

（3）搜索引擎法。搜索引擎是 Internet 上专门提供查询服务的一类工具，它利用被称做 Robot、Spider 或 Worm 等名称的自动化处理软件，定期或不定期地在网上爬行，通过访问网络中公开区域的每一个站点，对富媒体信息资源进行收集、然后利用索引软件对信息进行自动标引，创建一个详尽的、可供用户进一步按关键词查询的 Web 页索引数据库。这种数据库的内容一般有标题、摘要或简短描述、关键词和 URL、文件大小、语种及词出现的频率和位置等。搜索引擎方法是目前 Internet 上对富媒体信息进行组织的主要方式之一，较著名搜索引擎有 Google、Baidu、Alt Vista、Excite、Webcrawler 和 Lycos 等。此种方式所收集的信息虽然丰富广博，但良莠不齐，因而查准率低。

（4）主题树法。主题树法是将富媒体信息按照某种事先确定的概念体系分门别类地逐层加以组织，用户先通过浏览的方式层层遍历，直到找到所需要的信息线索，再通过信息线索连接到相应的信息资源。许多著名的网络检索工具如 Yahoo!、Sohu 等都采用这种方式组织信息资源。主题树法提供了一个基于树浏览的简单、易用的网络信息检索与利用界面。信息检索由用户按照规定的范畴分类体系，逐级查看，按图索骥，目的性强，查准率高。采用树形目录结构组织信息资源，具有严密的系统性和良好的可扩充性。当然，该方法也存在一些缺点，其中最突出的是必须事先建立一套完整的范畴体系，主要源于富媒体信息资源的庞杂性及信息用户对体系的了解，增加了用户的智力负担。另外，为了保证主题树的可用性和结构的清晰性，范畴体系的类目及每个类目下的信息索引条目都不宜过多，这大大限制了一个主题树体系所能容纳的信息资源的数量。

（5）学科门户组织法。学科信息门户（SIG，Subject Information Gateway）是针对特定学科或主题领域，按照一定的信息选择和评价标准及规范的资源描述和组织体系，对具有一定学术价值的网络信息资源进行搜集、描述和组织，并提供浏览、检索和导航等基本功能。学科门户组织法是由专业人员搜集富媒体信息，对信息源进行质量鉴别，使用受控语言和关键词进行必要的内容描述，从学科属性、研究方向和资源类型角度进行多重揭示的富媒体信

息组织方法。学科门户组织法融合了学科链接导航、搜索引擎法、主题树法、语义信息组织法和 Web2.0 信息组织等技术与方法,是具有较高学术价值的综合性数据服务平台。

1.2.4 按信息的认识层次划分

从人类对信息的认识层次上看,由于人类具有感受力,能够感知事物运动状态及其变化方式,由此获得的信息为语法信息;由于人类具有理解力,能够领悟事物运动状态及其变化方式的逻辑含义,由此获得的信息为语义信息;又由于人类具有明确的目的性,能够判断事物运动状态及其变化方式的效用,由此获得的信息为语用信息。对这三种信息进行组织,便产生了信息组织的基本方法。

1. 语法信息组织

语言学中的语法学是研究语言符号之间的结构规则的,主要包括词典构成和变化、词组和句子的组织,研究的语言结构属于形式范畴。信息组织借用“语法”二字,按照形式特征组织信息。最常见的语法信息组织有如下四种。

1) 字顺法。

这种方法是历史最悠久、使用最广泛的一种信息组织方法。它的意义在于从字、词角度集中排列有关信息,满足用户检索的一般要求。具体操作有音序法、形序法和两法并用等三种形式。从字顺组织法的发展史来看,目前音序法逐渐占据了主导地位。

2) 代码法。

信息用代码集中排列,人们易于接受又便于管理,所以随着信息量的激增和信息多样化,这种方法也从无到有,日益显示出其重要性。代码一般使用拉丁字母和阿拉伯数字,如专利代码和商品代码。

3) 地序法。

获取有关地域方面的信息是用户需求的一个重要方面,因为地域本身就代表了某一方面的特殊信息。同时,跨地域的比较研究日益增多也促使地域信息需求的增多。地序组织法一般有文字法和图文法两种形式。

4) 时序法。

同地序法一样,获得历史信息和从历史角度获取信息历来是大众的普遍需求。时序法有广泛的应用范围,如编纂工具书、著书立说,甚至写一篇文章等,都会使用时序法。

2. 语义信息组织

语义学(Semantics)中,语义具有研究语言符号与它代表的对象之间的结构关系之义。语义信息组织是专门研究信息的内容特性的一种信息组织方法,最常见的语义信息组织法有两种。

1) 分类法。

人们的分类对象可概括为三种:一是实物,如商品;二是概念,如知识;三是概念与实物的结合体,如文献。于是有三种不同的分类方法。在信息组织的实践中,它们可以结合使用。知识分类法是人类认识客观世界的科学方法,对其他两种分类法有着指导意义。文献分类以知识分类为基础,结合了文献实体属性和信息利用的实际。实物分类带有更多的专业(或行业)特性和效用原则。

2) 主题法。

这种方法是字顺法在语义信息中的特殊应用。它既采纳了字顺法直截了当、便于检索的优点,又兼顾了相同内容聚集的特点,是人们从内容角度更直接获取信息的有效方法。最常

见的主题组织方法有标题法、单元词法、叙词法和关键词法等。

3. 语用信息组织

语用信息是一些借助语用学的特有含义来研究随环境与使用者的不同而不断变化的信息群。常见的语用信息组织法有以下两种。

1) 权重值信息组织法。

权重值信息组织法就是按照信息的重要性来组织信息。例如，报纸在版面安排上，总是把最重要的信息放在头版头条的位置；电视节目的安排，总是把重要节目放在黄金时间播出。城市规划、行业决策和质量评估等其他地方也常用到这种方法。

2) 概率组织法。

这种方法是在未全知信息情况下对信息进行组织的方法，它根据事件发生的概率大小序化信息，可以在一定程度上规避风险。如预测文体活动胜负、期货交易等。

1.2.5 按信息的存在环境划分

在我国信息化建设过程中，建立和逐步完善一个能够使信息资源得以充分开发和有效利用的信息环境具有重要意义。信息环境是指与人类信息获得有关的一切自然、社会和心理的因素的总和。信息环境主要指与人类信息活动有关的人的要素、信息的要素、技术的要素和社会诸要素（政治、经济、文化、科技及政策法律等）。可见，信息存在环境既包括技术环境，也包括社会环境。如从是否使用网络技术的角度，可以将信息划分为网络信息和非网络信息。而从信息存在的社会环境角度，信息组织类型可以有以下四种。

1. 政务信息组织

政务信息也称政府信息，是涉及国家政府各级部门和单位的一切制度及事务范畴的信息总和。政务信息能够为政府部门决策提供依据，在信息化和网络化高度发达的当今社会，具有举足轻重的意义。政务信息组织应以有利于政务信息的“公开、公正、及时、透明”原则为准，做好政务信息的选择、分析、描述、揭示和存储等工作。目前，电子政务成为政务信息组织的主要方式，电子政务是政府机构应用现代信息技术，将政府管理和服务通过网络技术进行集成，在互联网上实现政府组织结构和工作流程的优化重组，打破时间、空间与部门分隔的限制，全方位地向社会提供优质、规范、透明和符合国际水准的管理和服务。

2. 经济信息组织

经济信息是反映经济活动实况和特征的各种消息、情报、资料 and 指令等的统称。在宏观经济、中观经济和微观经济活动中，都存在大量经济信息，人们通过其接收、传递和处理经济信息，反映和沟通各方面经济情况的变化，借以调控和管理生产，实现管理环节间的联系。经济信息又分为计划信息、控制信息、生产和经营信息以及统计信息等。通常，经济信息的组织多是从市场调研与竞争情报等活动中开始的，也有部分表现为经济类文献数据库，如国研网等。经济信息组织以具体的经济利益为目标，对相应经济信息进行组织与处理，形成高经济价值的信息资源。

3. 科技信息组织

科技信息是记录科学技术发展的信息集合，能够集中反映各时代、各领域、各学科和各行业科学技术发展水平与方向，具有知识性、创新性、敏感性和机密性等特点。科技信息组织是以相应领域的科技信息数据库为主，各类信息数据库包括中外文文献数据库，涉及电子

图书、期刊、学位论文、会议文献、标准文献、专利文献、研究报告和技术说明书等形式，科技信息组织产品是目前信息组织的最重要实施对象。

4. 文化信息组织

随着我国公共文化服务政策的逐步出台以及国家对文化重视程度的不断加深，文化信息的组织与管理的重要性日益凸显。文化信息积淀在人类文化发展的整个历程中，记录了一个国家或民族的历史、地理、风土人情、传统习俗、生活方式、文学艺术、行为规范、思维方式和价值观念等内容。目前，系统性的文化信息组织工作在我国仍处于初级阶段，但随着国家、组织或个人对文化信息组织的日益重视和研究探索，文化信息组织在未来将为国家文明建设、公共文化服务提升带来不可估量的巨大能量。

当然，以上四种信息组织方式是依据信息所存在的社会环境来区分，仅具有相对意义。其他存在环境下的信息及其组织方式也是大量存在。需要说明的是，非网络环境和网络环境的信息及其组织方式在前面已有介绍，在此不再赘述。以上各种类型的信息组织往往是相互联系、相互交叉、相互渗透的，在一定条件下可以相互转化。

1.3 信息组织的作用

由于现代信息技术和互联网络的迅速发展，信息生产、再造、存储和传播空前便捷，信息量急剧增加，随之而来的是信息质量参差不齐、信息污染日益严重、信息垄断与不公，“信息爆炸”、“知识爆炸”成为现代社会的一大特征。这些现象造成了人类与信息之间的两种矛盾：一种是知识和信息的海量性和无限性与人的精力、时间的有限性的矛盾，另一种是知识和信息的无序性和污染性与人类使用的选择性的矛盾。

美国未来学家奈斯比特在《大趋势》一书中曾说过，“我们淹没在信息中，但是却渴求知识。”为什么当各类信息像洪水一样向我们涌来时，我们仍然缺乏所需要的信息呢？这是因为在知识经济和信息化社会中，“失去控制和无组织的信息不再是一种资源。”

具体说来，信息组织对于信息交流与共享、重组与增值有如下作用。

1.3.1 控制整序作用

随着人类文明的产生和人类对世界的探索，信息资源数量不断增加。尤其是 20 世纪 90 年代以来，在现代信息技术和互联网发展的带动下，信息数量指数级不断增长，信息的混乱程度日益加剧，导致用户获取所需信息的成本越来越大。因为缺乏控制，信息集合不可能为用户提供高效快速的检索，信息数量越大，混乱程度越高，当这种混乱达到一定程度后，信息检索几乎无法进行。因此无论手工检索和计算机检索都离不开信息组织对信息的控制和整序，网络环境下的信息组织的作用就更为突出。当前，人们开发研制的各种网络信息组织和检索工具，对于网络信息的快速有效查询具有重要意义。

1.3.2 提升品质作用

随着信息数量的急剧膨胀，信息质量日益恶化，大量的无用信息、虚假信息和垃圾信息充斥于各类信息交流渠道，极大地妨碍了人们对有用信息、正确信息的吸收和利用。事实上，没有组织或不加控制的信息不仅不能创造财富，而且会对人们利用信息构成严重的危害。信息组织的作用之一就是通过对海量信息的选择、评价、整理、集成和重组等一系列工作环节，

提升信息品质，把良莠不齐的信息加工成为精良有序和高度集成的信息检索系统，为信息增值（如信息的分析研究）打下基础。

1.3.3 传播利用作用

信息组织克服了混乱的信息流带来的信息查询和利用的困难，把分散的、无序的信息加工为系统的、有序的信息流集合，使其成为真正意义上的信息资源，一方面便于信息机构通过各种方式向用户提供信息服务，另一方面也便于用户自主地、随时随地地查找和利用信息，从而充分发挥信息的效用。信息的交流实质上是信息传播的一种方式，或者说是信息传播的结果。序化的信息有利于信息的传播，研究表明，按一定规律排列的信息更易于人们理解和记忆固化，因而传播效能会更大。例如，报纸上的广告经过分类处理后，由于查找方便，读者接触到它的可能性明显增加。

1.3.4 节约成本作用

当前，信息获取与筛选的成本日益增大，用户的时间成本将大于信息管理和信息服务的价值。如果任其发展下去，将大大增加用户社会各项活动的总成本。信息组织将多余的和虚假的信息过滤掉，将有用的信息精选出来并加以序化，使得信息交流和存取等环节的效率不断提高，成本不断下降，从而节约了社会各项活动的总成本，使社会在同样的时间里能够创造出更多的财富。

1.4 信息组织的发展

作为信息活动的重要组成部分，信息组织是与人类社会同时产生并同步发展的。信息组织活动具有悠久的历史，在当今网络信息时代对信息组织需求的有力推动下，其理论研究方兴未艾，其实践工作不断丰富。

信息组织的发展过程是信息管理整体发展过程的组成部分，在发展阶段上，信息组织的发展大致与信息管理发展基本保持一致。由于在原始信息管理时期几乎谈不上任何形式的信息组织活动，信息的传播通常是不正规的、自发性的。因此，在此仅把信息组织的发展区分为三个阶段，即古代信息管理时期、近代信息管理时期和现代信息管理时期。需要指出的是，信息组织发展阶段的演化并不是一个阶段对另一个阶段的全面否定，而是一种扬弃、一种理论与方法的完善和扩展。即使是最早的清册职能也并未在现代信息管理时期被彻底废止，而是被融到一种更为综合的组织活动中。

人类的信息组织活动源远流长。了解信息组织的历史渊源和发展沿革，可以使我们从更深远的角度认识今天的信息组织活动，集成和汲取先人在长期信息组织实践中积累的有益经验。同时可以使我们高瞻远瞩、明确方向，使信息组织的理论和实践在科学地解决现实问题的基础上具有前瞻性，适应不断发展的社会需要。

1.4.1 古代的信息组织

古代信息组织时期是指近代工业革命之前的漫长的历史时期。信息组织最早是在人的思维认识领域中进行的，可以说自从有了人类大脑的思维活动，就有了最原始的信息组织。随着人类社会活动的日益复杂化和多样化，人类大脑的思维、分析、推理、归纳、类比和聚类

等功能得到不断加强,人们能在纷纭的信息中选择,将同类信息加以集中,不同的信息加以区分,而且能够由此及彼、由表及里地发现事物间的内在联系。人的大脑是在对信息的分析、综合、比较、抽象、概括、类比、判断和推理等活动中形成了最原始的信息组织方法。正如美国学者沃尔曼(R. S. Wurman)所比喻的:“人完完全全是由存储器组成的,这种存储器需要具有一个组织框架,该组织框架来源于个人的洞察力——具有精密过滤镜的眼睛、拥有独特频率响应能力的耳朵,以及通过思考找到个人兴趣的内容间相互联系的大脑”。尽管后来许多信息组织的活动和研究远远超越了人的认识领域而延伸至社会信息的生产、加工处理、存储、传播和检索的广阔空间,但一切信息组织方法都离不开人脑思维活动这个基础,都是在此基础上发展起来的。

信息符号作为表达概念的标识、信息组织的工具,它的产生和广泛运用对信息组织的发展起到了关键作用。人们对信息的内容和形式的组织是通过对信息符号的组织完成的。人类最早用手势、体姿符号来表达和传递信息,后来出现了有声语言符号,它比体姿符号能表达更多的意义,不同的语言表达不同的意义,这其中就包含了信息组织的过程。

为了克服体姿语言和有声语言不便保留、稍纵即逝的局限,我们的祖先很早就开始在寻求信息记录的形式,例如,曾采用“结绳”这一方法来记录事物的数量和特点:“事大,大结其绳;事小,小结其绳。”这是人类以实物作为符号来记录、描述事物的早期应用。以后,人们又开始用图画来表达思想、记录知识,并在此基础上产生了最初的文字。文字的产生和运用使信息可以脱离人脑而被记录、储存于外界物质载体上,使信息变得容易识别、存储、组织、传播与交流,是人类认识和处理信息能力的一大飞跃。

承载信息符号的各种载体是信息组织活动的物质基础,随着社会生产力的发展,信息记录的载体材料和信息记录方式不断改进。人们将代表某种意义的符号刻录在龟甲、金石上,书写在绵帛、竹木上,或者记录于泥版、羊皮上,进而书写在纸张上,刻录在光盘、磁带上,为信息组织的社会化创造了条件。

1. 中国古代信息组织的萌芽与发展

原始的信息组织还是自发的、不规范的信息组织。当人类的信息记录积累到一定数量,并被有意识地汇集、排序时,就开始出现了社会信息组织活动的萌芽。在我国的殷商时代,有了大量记录事物的甲骨文文献。据殷商考古发掘发现,窖穴中的甲骨文排列有一定顺序,甲骨上除了有卜辞记录之类的记录外,还刻有占卜者或经管卜官邸名字。这说明,殷代文献的管理就已经包含着目录工作的雏形,可以看做是我国古代目录工作的起源。在国外,最原始的“目录”是以泥版的形式出现的。约公元前 2200 年,在两河流域巴比伦王朝的一座寺庙的附近,有大批泥版文献被集中在一起,并按主题排序。这说明,当时对这些记录有信息的甲骨和泥版的收集和储藏是有意识的、经过整理的,当时已有了简单的标识著录法及排列法。

我国古代正式的信息组织活动主要体现于对文献的整理,通过对“部次甲乙”来满足人们即类求书、因书究学的需要。

最早的文献整理可以追溯到春秋时期孔子整理“六经”及以后各种图书目录的编制。孔子做学问时有感于大量文献的零散无序、不便使用,于是对鲁、周、宋和杞等故国的文献进行了收集、选择、校订和分类编排,将虽已成书但分散混乱的文献重新加以组织,使之成为结构合理的有序整体。如对《诗经》305 篇的组织整理、对《尚书》按体裁的分类排序,开创了我国私人整理文献的先河。

汉代刘向、刘歆父子编制我国第一部大型分类目录《七略》,解释了当时国家收藏的自先秦以来所有的图书,这是我国历史上第一次系统的、大规模的文献信息组织活动。刘氏父子将当时收集的所有藏书分门别类地著录,按学科内容分为六艺略、诸子略、诗赋略、兵书

略、算术略和方技略六大类，加上“辑略”部分共七类，形成了我国第一部综合性的国家书目。此后，各代又有多种有影响的目录出现，如魏晋南北朝、唐宋元明所编纂的官修目录、史志目录、佛经目录等。南朝王俭的《七志》、梁朝阮孝绪的《七录》采用了七分法，郑樵在《通志·艺文略》中创造了独具一格的12大类三级类目的分类体系，在我国使用最久、影响最大、占主导地位的是四分法。西晋时，国家书目《中经新簿》最早提出四部分类方式，后经逐步改易，至唐魏征所编的《隋书·经籍志》确立了经、史、子、集四部分类系统。自唐以来，各种官私书目大多采用了四部分类。清代《四库全书总目》收录了先秦至清初重要书籍10 254种，著录书名、卷数、著者书籍来源并有内容提要，组织成经、史、子、集四部44类，代表了我国古代目录的最高成就。这些目录系统揭示了我国历史文化典籍及其所记录的知识，为存储、查找信息提供了方便。这时的信息组织活动已经从单纯的个人行为发展到既有个人又有国家组织的社会活动，其范围和影响力不断扩大。

在主要应用分类组织法的同时，字顺主题法和索引法也初露端倪。《四库全书总目提要》中将颜真卿看做是采用声韵字顺的方法来排列事物主题的鼻祖，称“昔颜真卿编《韵海镜源》为以韵隶事之祖。”这说明，我国在唐代就已经出现了以事物或对象名称做标题的信息组织方法。

类书被视为我国主题法的滥觞。目录学家姚名达认为，类书可以视为主题目录的扩大，如果删其繁文，仅存书目，也就成为现代最先进的主题目录了。我国最大的类书《永乐大典》就是一部带有主题法性质的检索工具。全书22 877卷，采用了“用韵以统字，用字以统事”的组织编排方式，与主题法的字顺系统原则完全吻合。

中国古代索引是在字书、韵书和类书的基础上发展起来的，曾被称为“通检”、“备检”或“串珠”等，形象地说明了它的作用和特征。早在魏建安年间类书《皇览》就具索引功能，被认为是中国古代索引的起源之一。我国唐代林保的《元和姓纂》、宋代黄邦先的《群史姓纂韵谱》、陈思的《小字录》等即为早期的姓名索引，宋、明、清等时期还编纂了一些专书索引。特别是清代史学家、目录学家章学诚在其《校雠通义》等书中明确提出了一系列较为重要的索引理论和索引方法。

在实践的基础上，我国古代一些学者对信息组织的原理和方法进行了理论探讨和阐述。春秋末期，一些大思想家已经注意到事物分类的普遍性，从“类”的角度去认识事物和进行推理。如墨子不仅将“类”作为分析事物的依据，还把“类”与“故”（理由）联系起来作为立辞和辩说的基本原则。宋代目录学家郑樵从理论上研究了文献的分类问题，提出了一系列类分图书的原则和方法。他在《通志·校雠略》中指出：“类书，犹持军也。若有条理，所多而治；若无条理，虽寡而纷”，“类例不明，图书失纪”，“类例既分，学术自明”。生动阐明了文献分类的意义。清代著名学者、目录学家章学诚则进一步主张文献分类应随文献增长的变化和学术演变而改变，主张编制目录不仅要文献分门别类地加以编排，而且要“辨章学术，考镜源流”。他大力提倡采用“互著”、“别裁”（相当于现在的参照法和分析著录法）的方法对文献进行著录，并阐述了索引的方法和功用，主张将各种文献中的人名、地名、官阶、书名等按韵编排，详著出处，以便分类时按韵查找。阐述这些思想的专著《校雠通义》成为我国古典目录学理论的集大成之作。

2. 国外古代信息组织的萌芽与发展

在古希腊，一些哲学家很早就提出了知识分类思想，并构筑了自己的知识分类体系。例如，亚里士多德将知识区分为理论知识（逻辑学、物理学、数学、形而上学）、实践哲学（伦理学、经济学、政治学）、创造哲学（史学、修辞学、艺术）三大门类，对后世分类法中学科的划分产生了深远影响。中世纪的欧洲修道院图书馆和大学图书馆则分别建立了以宗教观

点和学科设置为基础的分类体系。印度的一些古老经文如《吠陀经》将人类全部知识分为四大门类,也包括许多朴素的分类思想。

古代国外的正式分类活动主要可以追溯到公元前6世纪。当时亚述巴尼拔皇宫收集了大量的泥版文献,上刻主题标记,并被分为人间、天上两大类,其下再分子类。公元前250年左右,古希腊学者卡利马科斯(Callimachus)为当时规模最大的亚历山大图书馆编制了长达120卷的藏书目录《皮纳克斯》(Pinakes,又名《各科著名学者及其著作目录》),将藏书分为戏剧、诗歌、法律、哲学、历史、修辞学、医学和杂著等大类,其下再按字母或年代顺序排列,并附对每部著作的评价,成为古代最早的目录之一。

被称为“目录学之父”的德籍瑞士学者格斯纳(C. Gesner)将当时能搜集到的全部拉丁语、希腊语和希伯来语的各科著作1.5万种编织成大型书目——《世界书目》,全书四卷,包括著者字顺目录、分类目录和主题字顺索引,其中,第二卷将知识分为21个大类、250个细目,较全面地反映了当时的科学发展水平,成为西方第一部检索系统较为完备、著录详尽的综合性大型书目。西方十三四世纪的《圣经》语词索引直接以自然语言中的字、词作为标目,按字顺次序查检,基本具备了主题法的要素。虽然当时主题和索引方法应用很少,但它成为近现代主题法和索引发展的渊源。

公元16世纪前后,随着罗马帝国的衰亡,欧洲社会逐步进入宗教统治时期,宗教文献及其记录与检索宗教文献的工具——目录、索引、文摘在西方国家有了迅速的发展。17世纪末,西方国家不仅出现了综合性与专题性的信息目录,而且在英国和德国诞生了具有现代意义的国家书目。1605年问世的《英国牛津大学图书馆目录》首次提出了基本款目的概念,为信息组织的理论探讨迈出了第一步。古代信息组织的主要活动是书目工作,这是一种对现有著述或文献特征进行识别、选择、著录与排序的信息整序工作。书目工作主要研究“已知”文献,对原始信息记录具有较强的依赖性。

总的来说,随着古代文明的发展,人类的文献生产、制作、积累与利用逐步形成规模,面对数量日益增长的文献,如果不加以组织与管理,人类将无法有效利用。基于这一点,早期的信息组织为适应人们管理与利用文献的需要而诞生。可以说,一定数量文献的生产与积累是信息组织产生的基础,而大量文献的产生与积累则是信息组织发展的首要条件。没有一定数量文献的产生及文献资源的积累,就失去了信息组织产生、发展的物质基础。

3. 古代信息组织的特点

从总体上看,由于古代社会生产力低下,信息的数量和对信息的需求都非常有限,在此基础上进行的信息组织活动基本处于初始阶段,呈现出以下特征:

(1) 以手工编撰书目的信息组织形式为主流,包括一书目录和群书目录、国家书目、联合目录等,同时也出现了以数序、音韵、类名和语词为排列方式的类书及从文献正文中摘录只语片断的文摘方法。

(2) 信息组织活动以个体劳动为主,虽然偶尔也存在着简单的协作关系,但并未形成一种长期固定的专门机构,如我国古代《七略》、《七志》的编撰大都以个体劳动为主。

(3) 信息组织着眼于文献的收藏管理,主要是为了管理一定范围内的文献,并非着眼于文献检索和利用;其信息组织的主要方式限于记录与登载文献的基本特征,或者按照学科或事物范畴排序;其成果形式类似于文献清册,即文献书目。

(4) 信息组织的主要对象是图书和档案文献。

(5) 信息排序多采用分门别类的方式,国外偶尔也采用主题(Catchword)方式揭示和排序等。总的来说,不管古代的信息组织其形式与成果如何简单,但其“辨章学术,考镜源流”的功能及分类揭示与排序的手段对后来信息组织具有深刻的影响。

1.4.2 近代的信息组织

近代信息组织时期是指从近代工业革命和印刷工业机械化到第二次世界大战（20 世纪 40 年代中期）的一段时间。此时期信息组织的总的特点是文献生产、积累与利用形成了较大的规模，信息组织的职能工作的主体由个体逐步趋向社会专门机构（如图书馆等）。

1. 国外近代信息组织的发展（17 世纪起）

近代工业革命极大地促进了社会信息活动的发展。以蒸汽机为核心的动力技术与活字印刷术的结合，进一步提高了文献生产的效率；交通运输工具的进步既密切了世界各国的联系，又为文献资料的传播提供了便利条件；以电力技术为基础的电信技术则为人类信息交流创造了新的手段。与此同时，近代科学的发展也为近代信息管理和信息组织活动开辟了广阔的舞台，提供了丰富的内容。科学研究活动从科学家个人的自发研究成长为有组织的社会事业，科学交流从自发组织的科学团体发展成为正规的学术管理机构，使科学劳动的成果成倍增加，文献信息的数量和需求也急剧增长。因此，图书馆作为社会上最早出现的信息管理和组织机构，在这一时期获得了很快的发展，文献组织的基本方法，如编目法、分类法和文摘索引法等，成为这一时期信息组织的主要方法。

近代以来，西方学术逐步进入发展时期，按研究对象划分的一门门自然科学开始从包罗万象的哲学中分离出来，人们的知识结构和思维结构发生了很大变化。新科学的产生和技术的发展在图书分类中得到了反映，出现了表达动力学、古生物学、细菌学、人类学、胚胎学、生态学和社会学等学科新概念的大类、小类和细目，出现了以学术发展为基础的知识分类体系。近代西方一些哲学家，如培根、康德、黑格尔、霍布斯和孔德等构造的不同知识分类体系对信息分类的思想产生了深远影响。其中较具代表性的是培根的分类思想和知识分类体系。他从心理特征出发，将知识分为历史（记忆知识）、诗歌（想象知识）和哲学（理想知识）三类，并在这三类之下又分出细类。这一分类体系后来被用于 18 世纪狄德罗编制的法国百科全书的组织，不少图书馆和藏书家也根据它来编制图书分类目录。恩格斯早在 1858 年就注意到了科学分类问题，他认为：“每一门科学都是分析某一个别的运动形式或这一系列互相关联和互相转化的运动形式的，因此，科学分类就是这些运动形式本身依据其内部所固有的次序的分类和排列。”

随着学科的不断变化、知识记录的不断增多、大型图书馆藏书量的急剧增加，专业图书馆也得到了发展。在这种情况下，原有的分类法已不适用。在旧分类法中没有包含的、由于科学和技术的发展而出现的新的知识门类的大量涌现，不仅需要确定新知识及新学科在图书分类表中的位置，而且需要深入研究，使类表进一步扩展到图书分类结构，以便对几十万、几百万册图书及藏书目录进行组织。

1876 年，美国图书馆学家、教育家杜威编制了《杜威十进分类法》（DDC）。这部分类法建立了结构完备、等级分明的分类体系和主题索引，体现了当时信息组织的最高水平。这一时期的分类法还对各种分类技术，如标记符号、类目索引等进行了探索，努力使类目体系具有“伸缩性”、“适应性”，类目符号具有简易性和助记性，出现了一批有影响力的综合性分类法，如克特的《展开式分类法》、奥特勒和拉封丹的《国际十进分类法》（UDC）、《美国国会图书馆分类法》（LCC）等。在杜威之后，编制统一的国际分类法的思想已经成了各国图书馆员关注的问题。

人们在对大量文献检索的过程中，愈来愈感到传统的分类法已不能完全满足需要。读者除了对分类的族性检索要求外，还开始关注对具体主题的特性检索。于是一些专家对以满足特性检索为目的的主题法进行了进一步探讨。18 世纪后，出现了主词索引和字顺分类目录。

主词索引直接在书名中抽取自然语言形式的关键词作为文献主题标目,成为后来标题目录的先导。字顺分类目录将字顺方法与分类体系相结合,在每一级类目下以字顺方式排列同类类,并逐步演变为将类目直接按主题字顺方式编排,也为主题目录的发展创造了条件。

主题法的真正形成和发展是在19世纪后半期。最早期的类型是传统的标题法。1856年,英国克里斯塔多罗的《图书馆编目技术》提出应用文献题名中的“词”作为字顺标题系统中表达内容主题的标题,成为标题法的先导。在这部书中,他提出了“主词”(即关键词)这一概念,并介绍了用轮排方法编制的曼彻斯特公共图书馆目录书名语词索引的步骤。1876年,美国图书馆学家克特发表了《字典式目录条例》,该条例在传统主词款目和字顺分类目录的基础上,明确规定了标题的意义和处理方式,制定了标题选择和使用的一系列原则和方法,从而完成了字顺分类法向字顺主题法的转变,它标志着现代主题法原则的确立。1895年美国出版了第一部标题表——《美国图书馆协会标题表》并在其后出版《美国国会图书馆标题表》(LCSH)。在19世纪后10年,主题目录在美国公共图书馆获得了广泛应用,引起了许多欧洲分类理论专家的关注。

同时,更具信息报道和指引意义的文摘和索引在此时期也发展起来。1830年,世界上第一部科技文摘杂志《药学总览》在德国问世,并附有索引。此后,各国相继出版了一系列文摘杂志和著者、题名、语词、分类索引。如1848年美国普尔的《普尔期刊文摘索引》、1851年美国的《纽约时报索引》。为开展有关研究和交流,英国于1856年成立了世界上第一个索引学会。

2. 中国近代信息组织的发展(鸦片战争后)

在漫长的封建社会发展过程中,我国的信息组织活动一直以藏书楼的建设为主体,与其相适应,四部分类法具有稳固的主导地位。鸦片战争以后,由于西学输入、社会变革和教育的倡导,反映新学科、新技术、新思想、新文化的信息内容日益增多,我国旧有的四部分类法已不能适应。一些学者开始以新的思想和方法编制目录,打破了四部分类法一统天下的局面。一批反映西学、新学的书目问世。例如,梁启超的《西学书目表》就摒弃了传统的四部分类法而采用了自编的新分类法体系。其中,将所收入的西学翻译书籍分为学、政、杂三大类23小类,初步具备了现代分类法将知识划分为自然科学、社会科学、综合性图书三大部类的雏形。1904年刊印的《古越藏书楼书目》也完全打破了四部分类法的传统,将浙江绍兴古越藏书楼的所有藏书分为学、政两大部,容纳了反映新学科的书籍。

19世纪末20世纪初,一些学者开始翻译、引进国外分类法。1909年孙毓修在《教育杂志》上连载《图书馆》一文,介绍《杜威十进分类法》,引起了广泛关注。一批留学国外学习图书馆学的学者归国后,不但积极从国外引进新的分类法及其理论,而且着手对我国旧有分类法进行改造,编制了一系列中西合璧的新型分类法,以寻求解决我国文献分类问题。至20世纪上半叶,我国引进的国外分类法有23种,包括著名的《杜威十进分类法》(DDC)、《美国国会图书馆分类法》(LCC)、《国际十进分类法》(UDC)、《克特展开制分类法》(EC)等。这对我国的信息组织体系结构和方法有相当大的影响。在学习研究国外分类法的基础上,我国相继出现了仿杜、补杜及自编的各种分类法达三四十种。比较有影响的如1917年沈祖荣、胡庆生编制的我国第一部以标记符号代表类目的《仿杜威书目十类法》、1929年刘国钧自编的《中国图书分类法》、皮高品以杜威法为基础改编的《中国十进分类法及索引》等。在此期间,虽然反映旧学的四部分类体系仍在继续起着作用,但从总体上看,这个时期我国文献分类法的发展过程是一个引进与继承、融合与创新的过程。

在信息分类的实践活动不断发展的同时,我国图书馆学家在理论上对分类标准、分类原则、分类方法等也进行了深入探讨。如1920年之嘉在“科学分类之历史”中论述了科学

分类史及一系列国外学者的分类思想和科学分类体系，杜定友、查修等人系统研究了我国历代书目及图书分类体系，戴志骞首先提出了图书分类法的编制原则。1929年1月，中华图书馆协会在第一次年会上通过了编制标准分类法的四项原则：“中西分类一致；以创造为原则；以分类标记须易写易记，易识易明；须合中国图书情形。”这反映了当时的基本分类思想。

在主题法方面，虽然20世纪30~40年代我国已有少数标题表问世，如何多源1939年的《标题表》、吕绍虞1935年的《中文标题总录初稿》，但当时没有得到实际使用。

20世纪初，西方现代索引技术传入我国，30~40年代，我国兴起了一个编纂索引的高潮。如哈佛—燕京学社引得编纂处洪业等主持编制了一系列索引。还有叶圣陶的《十三经索引》、王重民等的《清代文集篇目索引》和开明书店编纂出版的《二十五史入名索引》等，并有《索引与索引法》和《引得说》等索引研究著作出版。

3. 近代信息组织的特点

(1) 信息产品进一步丰富，信息组织与管理作为一种相对独立的职业得以发展。这一时期图书馆、档案馆和情报部门等文献机构普遍开展了文献信息组织活动，并开始面向社会公众提供信息服务。国内外出现了一批致力于研究信息组织的专家。

(2) 用户需求成为信息组织关注的问题，“用户中心论”在信息描述、揭示与检索点形成与排序过程中得到认同与贯彻。信息组织的活动除围绕着文献的保存开展外，开始注重根据社会需要提供信息服务，开始重视对信息外在特征和内容属性的全面描述、揭示，并为用户提供更多的信息检索点。

(3) 文献分类法的建立和完善。1876年，美国图书馆学家、教育家杜威编制的《杜威十进分类法》(DDC)建立了结构完备、等级分明的分类体系和主题索引，体现了当时信息组织的最高水平。1933年，印度图书馆学家阮冈纳赞编制了世界上第一部分面组配式分类法——《冒号分类法》，并系统提出了分面分类理论。对现代分类法和主题法的发展产生了巨大影响。在此过程中，我国的分类法在引进国外分类法的基础上推陈出新，实现了从四部分类法向以科学分类为基础的体系分类法的转变，并结合我国文献的特点，力求解决中外文献、古今文献的统一分类问题。

(4) 信息组织技术与方法的逐步完善，信息组织的对象从载体单元逐渐转向知识单元，索引和文摘组织法的面世打破了书目形式长期一统天下的局面，对以满足特性检索为目的的主题法进行了进一步探讨。1830年，世界上第一部附有索引的文摘杂志《药总览》在德国问世。此后，各国相继出版了一系列文摘杂志和著者、题名、语词、分类索引。1848年美国普尔的《普尔期刊文摘索引》，1851年美国的《纽约时报索引》。

(5) 主题组织法和机械化信息组织手段的发展。1895年出版了第一部标题表——《美国图书馆协会标题表》(LCSH)，20世纪50年代美国的陶伯(M. Taube)以字面上不能再分的词汇单元——元词作为标识，以字面组配表达文献主题，并结合比孔卡等设备的使用，开创了在检索阶段匹配检索的后组式检索方式。它标志着继标题法之后，一种新的主题法——单元词法问世。这种方法的基本原理在美国海军兵器中心、美国专利局、杜邦公司等单位得到应用，为后来发展的计算机自动检索开辟了道路。

(6) 信息描述与揭示的内容日臻完善，信息存取系统的检索途径增多。如美国《化学文摘》(CA)，除了有著者索引和主题索引以外，还增加了环系索引、分子式索引和专利号索引等。

近代信息组织的发展与变化为现代信息组织的发展与完善奠定了基础。

1.4.3 现代的信息组织

现代信息组织时期是指以电子计算机技术、通信技术和网络技术等一系列现代信息技术的诞生、发展及在信息组织中广泛应用为主要特征的时期,即从20世纪50年代至今。20世纪中期以来,随着科学技术的飞速发展,人类生产信息、积累和利用信息的活动形成了空前的规模,出现了“知识爆炸”、新技术层出不穷的现象。以电子计算机、网络技术和多媒体技术为主的现代信息技术的发展不仅把信息组织推上了一个新的发展水平,也将信息组织的技术手段带入到一个新境界,信息组织不仅成为现代科学技术与国际经济发展的重要组成部分,而且已成为人类社会文明程度的重要标志之一。

大体上以20世纪90年代初期互联网的商业化为分界线,可以把现代信息组织分为基于计算机应用的信息组织和基于网络的数字信息组织两个阶段。

1. 基于计算机应用的信息组织阶段

这一阶段信息组织的发展的特点从宏观和微观两方面得以体现。

1) 宏观上的主要特点是信息组织的技术化、社会化、产业化和标准化

(1) 信息组织技术化。

自1946年世界上第一台电子计算机在美国诞生后,计算机技术便被应用于信息组织工作。如1954年,美国海军兵器试验站首次建立了世界上第一个计算机书目存取系统。同期卢恩(Luhn)首创了题内关键词索引,开创了计算机编制索引之先河。20世纪60年代以后,书目工作自动化有了进一步发展,首先是机读目录(MARC)研制成功。1966年美国国会图书馆成功地开发了机读目录MARC I,1968年又推出MARC II,并正式向国内外发行。机读目录的研制成功极大地推进了信息组织现代化的进程。20世纪70年代以来,信息组织的现代化领域不断扩展,从文献编目自动化走向文摘自动化、检索网络化、翻译自动化、索引自动化和自动抽词标引技术,其实现方案也层出不穷。例如,英国于1974年推出新型的索引系统——保持上下文索引系统,在世界上产生了很大反响。我国上海交通大学的王永成教授等人对自动文摘的研究已取得令人鼓舞的成绩。

另一方面,自20世纪80年代以来,除电子计算机以外的其他信息组织技术也发展很快。例如,1978年第一个光盘问世;1979年第一个实用光缆通信系统在美国芝加哥投入使用;20世纪70年代末80年代初,可视数据存取系统在英、法等国家的实际使用。所有这一切,都极大地促进了信息组织手段与技术的现代化。

(2) 信息组织社会化。

信息组织社会化起源于图书馆联机编目工作。早期的文献编目采用手抄、雕刻等手工方式,随后又转换为打印模式。20世纪50、60年代,计算机逐步应于编目领域,美国国会图书馆先后研制成功MARC I及MARC II后,采用电子及光电技术进行编目成为60年代后期崭新的计算机编目形式。MARC的优点是:可促进文献编目标准化;加速编目流程,加速书目信息的存储与传递速度;可一次输入,多次输出;可为编目社会化和网络化提供条件。60年代中期MARC II试制成功,标志着联机编目社会化和网络化的开始。美国是世界上实现联机编目最早的国家,其中最著名、影响最广的是联机计算机图书馆中心(OCLC)。到80年代中期,全美已有5000多个图书馆,约94%的新到资料利用这个联机系统中的书目记录进行编目,各馆还同时向该系统输入各自的原始目录记录,供其他单位合作使用。美国是社会化联机编目的先驱,为世界各国编目工作网络化提供了宝贵经验。

(3) 信息组织产业化。

信息组织活动产业化孕育于20世纪50年代,至70年代中期进入稳步发展时期。由于

信息组织产品可以转化为商品，大量文摘索引服务社和书目数据公司的产生与发展，使具有特殊社会经济价值的信息组织产品的生产和流通出现了产业化和商业化趋势，主要表现在两个方面。其一，信息组织的数据库产品逐步应用于各个领域，包括企业、政府机构、金融行业、科研单位、商业领域等，这就为其产业化提供了市场。数据库研制与服务由早期分散的、小规模形式逐步发展成为集中的、多数据库的大规模服务，并在全世界形成了几十个大型的跨国数据库服务中心，建成了国际性的数据库销售体系。其二，商用信息存取系统自 70 年代起，先后投入营运，如国际联机存取系统 DIALOG、STN 和 OCLC 等。与此同时，传统的印刷型文摘书目等检索刊物也纷纷转为磁带型、网络型书目数据库，如《化学文摘》(CA)、《生物学文摘》(BA) 和《科学文摘》(SA) 等均同时出版印刷型、缩微胶片型、磁带型与网络版。书目数据库信息组织产品开发的产值逐年增加。

(4) 信息组织标准化。

标准化是此阶段的重要特点之一。信息描述与著录经过 100 多年的实践，从随意编目到一国和跨国统一编目，直到 20 世纪才迈开国际化的步伐。其中，文献著录标准 ISBDs 计划自 1984 年基本完成后，很快就成为世界各国编制文献著录标准的主要蓝本。此外，西方国家，如英国、美国、加拿大等英语使用国的文献分类组织基本统一使用“LCC”和“DDC”，这两个分类法经过数十年的实际使用，屡次修订，已成为国际公认的较为成熟、质量较高、被世界各国广泛采用的文献分类体系。我国目前统一使用的《中国图书馆分类法》经过不断修订，其科学性和逻辑性也不断加强。不仅如此，文献主题标引工作标准化于 20 世纪 80 年代在西方国家普遍推广，除 LCSH 被广泛采用之外，分类主题一体化的模式也备受信息组织理论研究与实践领域的关注。我国的一体化分类主题词表《中国分类主题词表》也于 1994 年正式出版发行，加快了我国文献标引工作的标准化进程。在版编目(CIP)则是我国于 20 世纪 90 年代启动的一项信息组织标准化的重要措施。

2) 微观上的主要特点是对分类法的改造、主题法的发展、分类主题一体化、自然语言检索系统的兴起及对自动标引和分类的探讨

(1) 分类法的改造。

早期的分类法如杜威十进分类法、克特展开式分类法均属于列举式分类法。这种分类法的树型结构虽有结构明晰的特点，但不便揭示复杂、专指、细小的主题，也不便容纳新的学科主题。为此，分类学家们进行了多方面的改进。1933 年，印度的阮冈纳赞编制出世界上第一部分面组配式分类法——冒号分类法以后，其中所体现的分析—综合的分类理论引发了分类领域分面研究的热潮。在其影响下，一系列专业分面分类表问世。如英国分类法研究小组在深入探讨分面分类理论的基础上，编制了十多种专业的分面分类法，该小组成员奥斯汀还据此编制了保留上下文索引系统(PRE-CIS)。1955 年，该分类法研究小组向英国图协和联合国教科文组织提出：“需要以分面分类作为一切情报检索方法的基础”的著名备忘录，在 1957 年的国际分类研究研讨会上受到一致肯定，从而奠定了分面组配理论的基础地位。

同时，传统分类法被不断地增加分面组配成分，朝分面组配方向改造。如 DDC 的通用复分表从无到有，从 1 个增加到 7 个，专类复分和仿分也在增加，在第 20 版则用分面分类的方法对音乐类进行了全面的改造。1976 年，英国分类法研究小组的成员米尔斯(L.J.Mills)对美国图书馆学家布利斯 20 世纪 40 年代初编的《书目分类法》(BC1)进行了全面的分面改造，使其由原来的等级列举式分类法发展成一部大型的外面组配式分类法——《布利斯书目分类法》(BC2)，成为列举式分类法彻底分面改造的典范。

我国现代分类法的发展经历了从 20 世纪 20~30 年代模仿杜威法、50~60 年代模仿苏联分类法，到全面探索我国适用分类法的理论和方法的过程，初步形成了具有中国特色的分类

体系和分类理论。新中国成立以来,先后编制了多部大型综合性分类法和数十种专业分类表,最著名的如《中国人民大学图书馆图书分类法》、《中国科学院图书馆图书分类法》、《中国图书馆分类法》、《中国档案分类法》和《中国标准文献分类法》等。尤其是大型综合性分类表《中图法》,广泛吸取了国内外各种分类法的优点,以科学分类为基础,同时在四次修订过程中不断地扩大分面组配技术的使用范围,逐渐增加复分、仿分方法,引入并扩大冒号组配方法的应用,成为目前我国文献信息组织使用最广泛的标准化分类体系。

(2) 主题法的发展。

在陶伯(M. Taube)开创的单元词法问世以后。这种方法的基本原理在计算机检索系统得到应用。几乎与单元词法同时,1947—1950年间,美国的穆尔斯(C. N. Mooers)在研究组配分类法的基础上,提出了一种新型主题法——叙词法,并创造“叙词”、“叙词法”、“情报检索”、“情报检索系统”等专门术语。他认为,叙词是表达概念的一种标识符号,是十分自由和独立的观念成分,应能以任何组合或次序来规定一次检索。20世纪60年代,为适应计算机在图书馆及情报工作中的应用,叙词语言吸收了标题法、单元词法、关键词法及分类法等各种检索语言之长,逐步取代了元词法成为现代情报检索语言的主流。它以概念组配取代字面组配,并广泛揭示概念间关系,使文献信息的揭示更加准确。1959年美国杜邦公司编制了第一部叙词表。

在我国,1950年曾出版过程长源的《中文图书标题法》,填补了新中国中文标题语言的空白。20世纪70年代,我国开始大规模编制和使用主题法。1971年,航空部情报所编制的《航空科技资料主题表》第2版问世,成为我国的第一部叙词表。此后,《电子技术汉语主题词表》、《常规武器专业主题词表》、《国防科学技术主题词典》、《原子能科技资料主题词典》、《机械工程主题词表》、《汉语主题词表》陆续出版。尤其是1979年出版的《汉语主题词表》,它作为“748工程”(即汉字信息处理系统工程)配套项目,历时5年编成,不仅成为世界上最大规模的叙词表,而且为后来我国叙词语言的大发展做了人才和理论准备。

(3) 分类主题一体化。

以往分类法、主题法作为信息组织的两种独立的方法各成体系,在使用过程中,标引和检索都须分别进行。计算机在信息组织中的使用进一步揭示了分类系统与主题系统的联系,促成了分类主题一体化的发展。1969年,英国学者艾奇逊(J. Aitchison)成功研制出世界上第一部分分类主题一体化的《分面叙词表》,它将一部分面分类表与一部字顺叙词表结合起来,通过严格规范,使每一个词汇同时出现在分类表与叙词表中,实现了两种检索语言的兼容。在其影响下,英美等国陆续出版了一批分类主题一体化词表,如《伦敦教育分类法(第二版)》、《建筑工业叙词表》、《基础叙词表》等。此外,印度、日本、苏联、德国等世界许多国家都开展了分类主题一体化的研究和实践活动。

20世纪80年代,我国先后用手工和计算机编成《常规武器分面叙词表》和《教育分面叙词表》。这是我国图书情报界编制一体化词表的最早尝试。此后,我国又陆续编制、出版了十余部一体化词表,包括三部大型词表——《中国分类主题词表》、《农业科学叙词表》、《社会科学叙词表》和七部中型词表,使一体化语言理论与应用研究进一步深入。其中,《中国分类主题词表》作为国家哲学社会科学“七五”规划重点项目,对推进汉语分类主题一体化词表的发展起到了重大推动作用。

(4) 自然语言检索系统的兴起。

分类法和主题法采用规范化的人工语言系统控制文本信息,虽然规范严谨,但难以掌握,词表的稳定性强,对日新月异的新学科、新事物难以适应。因此,它们的使用长期被局限于科研和专业领域,很难成为大众信息检索工具。在新形势下,人们希望有更方便、更具亲和

力和大众性的语言系统以满足日益丰富的、动态的信息检索需求。

20 世纪 50 年代，卢恩在前人探索的基础上，将计算机用于关键词索引的编制。其后，各种直接以自然语言为标识的检索系统相继出现。20 世纪 60 年代初，美国匹兹堡大学健康法律中心率先建立起第一个全文检索系统，该系统以电子文本为处理对象，通过计算机以自然语言的语词或语词的组配进行检索，广泛探索了自由文本的检索方法，后来，图书馆界的书目检索系统吸纳了这一方法，将文本检索用于标题和文摘检索。

20 世纪 70 年代，一批采用句法—语义分析技术的自然语言理解系统脱颖而出，在语言分析的深度和难度方面都比早期的系统有了长足的进步。进入 20 世纪 80 年代后，适应知识管理的需要，智能信息检索系统开始应用于自然语言的处理。自然语言容易被普通检索者接受，但它也存在着表达概念过分自由、语义无关联、词汇无控制等问题，影响了检索效率。于是，一些专家提出了编制后控词表等方案以改进检索效果。

（5）对自动标引和分类的探讨。

自动标引始于 20 世纪 50 年代后期，1957 年，卢恩在对自动标引和自动编写文摘研究的基础上，发表了关于文献自动标引的论文，提出了基于词频统计的抽词标引法，率先进行了自动标引的探索。从 60 年代后期到 70 年代末，自动标引研究取得了很大进展，提出了概率统计标引法、句法分析标引法及各种加权模型等，建立了一批应用与实验系统。1962 年，博科（H. Borko）等人利用因子分析法进行了文献的自动分类。此后，各国学者进行了大量有关文本词频统计分析、句法分析、语义分析等实验研究，并在主要的国际联机检索系统中实现了全关键词自动标引和检索。主关键词标引也建立了许多实用系统。20 世纪 70 年代美国国防部文献保障中心（Defense Documentation Center）采用的机助标引系统、90 年代美国 NASA 宇航信息中心使用的机助赋词标引系统等，都是结合自动标引研究成功建立的人机结合的实用系统。20 世纪 80 年代末，日本庆应义塾大学文学系的图书情报专业和日本 IBM 东京基础研究所合作开发了一个自动分类专家系统，采用国际十进分类法（UDC）实现了图书资料的自动分类。1987 年，《杜威十进分类法》（DC）的联机编辑系统建成，并完成了对 DC 第 20 版的自动编辑。至今，自动标引形成了抽词标引和赋词标引两大主要类型。几乎所有大型分类法都实现了计算机辅助编辑修订和管理。

我国自 1980 年起开始从事汉语自动标引和分词实验研究，并逐步达到了科技文献自动分词的实用水平。20 世纪 80 年代中期起，我国在大量有关自动分类理论研究的基础上，相继开发出一批机辅分类标引系统、自动分类标引系统，如“计算机辅助分类主题标引系统”、“文书类档案的计算机分类标引算法”、“中文文本自动分类系统”、“中文技术文档的自动分类系统”、“图书分类专家系统”和“中文地质资料分类专家系统”等。

2. 基于网络的数字信息组织阶段

这一阶段的主要特点是：在以计算机、现代通信技术为代表的信息技术的推动下，传统的信息组织方法不断创新和完善，向自动化方向发展。同时，在网络环境下，信息组织面对多元化的信息载体和多元化的信息需求，又需要不断拓展新的应用领域，探讨新的数字信息组织方法。这两者之间存在着密切的关系。

1) 网络环境下传统信息组织方法的延伸和发展

（1）利用元数据对数字信息体进行描述。

元数据（Metadata）是关于数据的结构化的数据，专门用来描述数据的特征和属性。一个元数据款目构成一个信息资源的基本数据，是检索系统的基本构成单元，它可以代表信息资源用来组织目录、索引、数据库和搜索引擎等检索系统。信息描述的目的就是以元数据为中介，对信息资源进行各种操作。元数据的基本类型有：以详细记录为目的的元数据——机

读目录(MARC)、以发现为目的的元数据——都柏林核心(DC)、以网络查询为目的的元数据——搜索引擎。

MARC即机读目录,是以代码形式和特定结构记录在计算机存储介质(磁带、磁盘、光盘)上的,用计算机识别和阅读的目录。MARC作为一种以详细记录为目的的元数据,是电子化管理、发布类表、词表、人名、机构名等规范文档的标准方式。目前美国国会标题词表(LCSH)、国会图书馆分类法(LCC)都提供MARC版本并可以在网上查询。用MARC格式表示的知识组织系统可以植入联机公共目录查询系统(OPAC)中与书目数据统一管理。

都柏林核心元数据简称DC元数据,是为了描述数字信息资源、支持数字信息资源检索而建立的元数据模式。它是一种以资源发现为目的的元数据,旨在为数字信息的著录提供一种更简单、更有效的方式。它通过电子资源提供者对数字资源属性信息的描述,粗略地对资源内容进行编目,来帮助人们尽快地在网上发现所需要的资源。DC元数据避免了搜索引擎结构过于简单而MARC格式又过于复杂等问题,它不需要进行专业化训练就能对数字信息进行恰当的著录,降低了编目成本,能迅速适应数字信息资源的巨量增长,在数字信息组织中得到了广泛的应用。

元搜索引擎是一种以网络查询为目的的元数据,它使用自动索引软件来发现、收集并标引网页码,建立数据库;以Web形式提供给用户统一检索界面,供用户输入关键词、词组或短语等检索项;代替用户在数据库中查找出与提问匹配的记录并返回结果且按相关度排序输出。其特点是:由自动索引软件生成数据库,收录信息广、加工速度快,能及时向用户提供新增信息。元搜索引擎由检索请求提交机制、检索接口代理机制和检索结果显示机制三部分组成。检索时直接输入关键词或词组、短语,无须判断类目归属,使用比较方便。常用的搜索引擎有谷歌、百度等。但是,元搜索引擎依赖于数据库选择技术、文本选择技术、查询分派技术和结果总和技术,用户界面的改进、调用策略的完善、返回信息的整合以及最终检索结果的排序是未来元搜索引擎研究的重点。

(2) 虚拟图书馆(专题指引库)。

虚拟图书馆就是根据特定的目标,选定信息资源的学科领域,对有关的网站网页进行检索和收集,加以鉴定核实,并对核实后的网址进行合理组织,使之能够提供检索、浏览和链接的信息集合。与搜索引擎的主要区别在于:它属于专题性和学科专业性的,系统性和易用性强。虚拟图书馆对网络资源的组织是优越于搜索引擎的关键环节。该环节由专业图书馆员把关,在自动系统的协助下,利用某种分类法和主题词表,对收集来的原始资源进行描述和组织,改善了搜索引擎采用自然语言标引的根本缺点。这方面的实践和研究集中在重点学科导航库建设、专业网络资源导航库建设和热门站点链接或相关站点推荐方面。重点学科导航库是以学科为单元对互联网的相关学术资源进行搜集、评价、分类和组织的序化整理,并对其进行简要的内容揭示,建立分类目录式资源组织体系、动态链接学科资源数据库的检索平台并发布于网上,为用户提供网络学科信息检索线索的导航系统。如我国“高等教育文献保障体系”(CALIS)提出构建重点学科导航库系统,建议其内容可分为7项,并规定各子项目必须有分类浏览功能,以主题树浏览方式组织信息。

(3) 分类法在数字信息组织中的应用。

国内外学者就图书馆分类法特性、具体分类法与搜索引擎分类体系比较(类目涵盖范围、揭示深度、类表结构和功能)进行研究,认为传统分类法的知识系统性、标识语言的通用性、族性检索能力和扩检、缩检功能,是其他情报检索语言所不具备的。它在网络中的应用主要有如下几种。

其一,用于数字信息资源组织。以文献分类法为工具的网络资源检索服务系统,从学科

的角度揭示网络信息，成为组织网上学术性知识内容的主要应用模式。如欧洲科研与教育信息服务发展计划于1997年8月在“互联网资源描述与发现”专题报告中介绍了这方面的应用：有17个网上服务系统使用DDC组织资源、5个使用UC、5个使用LCC。国外在这方面的实验还有：使用DDC的万维网杜威分类法（CYBERDEWEY）、使用国会图书馆（LC）的万维网虚拟图书馆（WWW.Virtual Library）。研究表明，分类法能够提供网络信息资源框架，并作为网络环境下的浏览工具。

其二，建立网络信息资源分类法。现有众多的中文搜索引擎均采用各自分类体系组织网络资源，用户必须熟悉不同的分类体系才能较快地检索信息，这样给用户带来了极大不便。专家们提出，要建立网络信息分类法，提供网络资源的统一分类体系。其中，有的专家提出要建立网上信息的知识分类系统的基本结构与编制方法；也有专家提出以《中图法》为蓝本，综合国内外优秀分类法及搜索引擎的优点编制我国的数字化分类法；还有专家提出从最终用户检索需求出发改造《中国分类主题词表》。

其三，人工神经网络（ANN）。人工神经网络是根据人类的生物神经系统结构设计的计算机系统。在信息组织中可以用于自动分类。国外有些信息检索系统已经采用了使用ANN的自动分类系统。目前分类上应用最广泛的人工神经网络模型叫做自组织映射（SOM），利用该网络可以实现Web文档的自动聚类，如果在此基础上更进一步：即利用SOM网络实现索引词聚类，就可以实现超文本链接的自动生成。

（4）主题法在数字信息组织中的应用。

近年来，国内外学者及研究机构已认识到主题法在数字信息组织中的重要作用，在这方面的研究主要集中在以下几方面。

其一，关键词法的应用。由于关键词法具备：在标引时不必查表，选词、标引速度快，成本低；不依赖专职标引人员，可由作者或机器自动标引；不存在人为性或滞后性，能及时应用最新的提法及最新的词汇等优点。因此，目前由搜索引擎软件自动建立的网络信息资源索引数据库支持的就是关键词检索。但由于关键词检索的查准率低，所以人们提出使用后控词表（标引不控制+检索控制）的模式改进关键词法的检索性能。

其二，主题词表的应用。少数搜索引擎提供主题词检索方式，在用户界面上，可直接浏览主题词表，从中选取主题词，作为搜索引擎的检索提问。用户可以在检索界面中修改检索提问，也可返回到主题词表界面重新选择主题词。其共同特征是词表内具有超文本导航。

其三，标题词表的应用。标题词表在网络信息组织中的应用可以分为两种情况：一是检索前使用，即通过标题词表规范用户的检索表达式，用户可以首先在网络信息组织工具提供的词表中检索到标准的标题词及相关联的词汇，以该词作为检索词，单击表中超链接即可得到检索结果；二是检索后使用，即在给出用户所用检索表达式，得出检索结果的同时，提供相关词作为用户进一步检索的线索，用户可自由进行扩检和缩检，从而提高检索效率。

（5）主题图的应用。

主题图是一种新型的数字信息组织方法，使用这种方法可以提供最佳的信息资源导航。主题图利用了主题索引的概念及网站的特点，将主题、联系和范围三者紧密结合起来，用格式表单元来控制信息的获取和浏览，并详细描述各种浏览层次，实现对复杂知识管理关系的模拟，以便帮助用户更有效地浏览数字信息资源。由于主题图吸收了各种知识组织方法的长处，并采纳了本体（Ontology）和语义网的部分思想，它可对数字环境下的信息资源进行有效的组织与管理。主题具有良好的信息检索功能，表现在：可支持现有的搜索引擎在资源层面实现检索；主题图概念可看成一个图或树，支持可视化人机交互式检索；主题图可看成本体，提供一定程度的概念之间的关系描述，利用概念间的关系，提供一定程度的智能化检索。

(6) 本体的应用。

本体 (Ontology) 的概念源于哲学, 即对世界上客观存在物的系统描述, 一般译做本体论。许多研究人员从不同的问题和研究角度出发, 对于本体给出了不同的定义。Studer 等在对本体进行了深入研究之后, 给出了一个被广为接受的定义, 提出“本体是共享概念模型的形式化规范说明。”这个定义有 4 层含义: “概念化”指识别反映某些现象的相关概念的抽象模型; “明确”指所使用的概念及它们之间的联系都被明确定义; “形式化”指本体是计算机可读的; “共享”指本体中反映的知识是其使用者共同认同的。

从某种意义来讲, 本体同叙词表一样是一种控制词表, 是一种知识组织工具。虽然本体和叙词表同是知识组织工具, 但在形式化水平、概念抽象和语义关系表达等方面存在着明显差异。本体的结构特性使其在理论上具有超越叙词表性能的可能性, 并为它在一个更广阔的范围内获得有效应用奠定了基础。事实上, 本体的应用范围比叙词表更广泛, 信息组织与检索不过是它的一个适宜应用的领域而已。有学者从实现虚拟组织信息共享出发, 提出了一种基于本体的两阶段视图映射关系构建方法。以使得所构建的视图映射关系既能保证较高的信息查询效率, 又能保证具有良好的可扩展性。目前, 国内外的学者、专家正在积极研究如何基于叙词表来构建本体的问题。

2) 网络环境下新的数字信息组织方法与手段不断涌现

网络环境下新的数字信息组织方法与手段层出不穷, 下面从三大方面择要进行介绍。

(1) 从总体规划层面, 信息组织分为信息构建、信息网格、概念地图等方法与技术。

信息构建 (Information Architecture) 源于从建筑学视角来解决信息组织和利用的问题, 它通过合理地组织、标识信息并构建信息环境, 以改善信息浏览及信息检索的过程与效果的科学和艺术。信息构建的核心理念是关注用户、以人为本。信息构建是信息用户、信息内容与信息组织三者的交集。在对信息组织、导航、标识和检索各个环节的设计实施过程中, 应当时刻关注用户需求、用户习惯和用户利益; 可视性与可描述性也是信息构建关注的焦点, 它要求深刻理解信息环境, 明确信息在哪里、这些信息属于谁、如何利用这些信息及组织信息的目标是什么等问题。信息构建的主要活动是组织信息内容、生成信息结构和设计信息界面, 其直接目标是建立一个清晰的、易于理解的信息结构, 最终目标是提供给信息用户一个良好的信息空间环境。

网格技术是近年来国际上兴起的一种信息技术, 是互联网信息技术发展的新趋势, 包括计算网格、知识网格和信息网格等。将网络引入信息管理领域, 便产生了所谓的信息网格。信息网格就是要利用现有的网络基础设施、协议规范、Web 和数据库技术, 为用户提供一体化的智能信息平台, 其目标是创建一种架构在 OS 和 Web 之上的基于 Internet 的新一代信息平台 and 软件基础设施。在这个平台上, 信息的处理是分布式、协作式和智能化的, 用户可以通过单一入口访问所有信息。信息网格追求的最终目标是能够做到服务点播 (Service On Demand) 和一步到位的服务 (One Click Is Enough)。目前, 对信息网格的研究主要集中在体系结构、信息表示和元数据、信息连通和一致性、安全性等几个方面。

概念地图是针对特定主题的个人结构化知识的一种图示方法, 也是语义网络的可视化表示方法。概念地图用节点表示概念, 用连接线和连接词表示概念之间的关系。概念地图具有三个特征: 一是命题, 由两个以上的概念及其关系构成表达意义的陈述; 二是等级结构, 按照宽泛概念在上、具体概念在下的顺序排列形成等级结构; 三是交叉关系, 不同分支中的概念之间形成的连接关系。概念地图的构建包括四个步骤: 第一, 概念选取, 即列出关于某个主题的所有重要概念; 第二, 概念分类; 第三, 定位中心概念; 第四, 连接交叉概念。概念地图的构建过程即是知识创新的过程, 利用概念地图可以沉淀隐性知识。概念地图的构建过

程也是学习的过程，可以将学习中涉及的资源链接到概念地图，实现知识结构与相关资源的整合；同时，概念地图表示的知识结构遵循人类的认知和学习过程，因此可以利用概念地图导航用户检索所需的显性知识。

(2) 按信息服务方式，信息组织分为智能检索、个性化服务、信息可视化等方法与技术。

智能检索以文献和检索词的相关度为基础，综合考查文献的重要性等指标，对检索结果进行排序，以提供更高的检索效率。智能检索的结果排序同时考虑相关性和重要性，相关性采用各字段加权混合索引，相关性分析更准确；重要性指通过对文献来源的权威性分析和引用关系分析等实现对文献质量的评价，这样的结果排序更加准确，更能将与用户愿望最相关的文献排到最前面，提高检索效率。

个性化服务是指为了方便用户利用各种数字化资源，充分利用各种智能化技术对不同类型、不同特点的数字化资源进行整合，实现信息资源、信息技术、信息内容的集成，使目前信息资源组织系统的公共用户界面变得简单、友好，并且使用户能利用同一检索表达式或检索词对各种数字化资源进行同步检索，实现同一主题信息资源的一步到位的检索与查询的信息组织方法。同时，它还可以根据某些用户特定的信息需求定制具有个性化特点的用户界面，来提供符合其特定需求的具有个性化特点的信息和信息服务。个性化的信息服务是以信息资源整合和信息服务集成系统的建立为基础的，如果没有完备的资源整合体系作为后盾，无缝的、贴切的、高效的、主动的和一站式的信息服务模式则是无法实现的。

信息可视化利用计算机支撑的、交互的对抽象数据的可视表示来增强人们对这些抽象信息的认知。信息可视化的过程就是从信息维映射到可视维的过程。一般来说，通用信息的可视化可分四个步骤：抽取、转换、映射（定义）、隐喻。当今信息时代，人们常用“信息爆炸”来形容信息量猛增的特征与趋势。要处理和应用这浩如烟海的信息，需要用先进的处理方法和有效的工具。信息可视化就是将信息转换成二维和三维图形、图像、动画形式的技术方法和有效工具。用户通过这些可视形式进行观察、交互。图 1.5 为《中图法》22 个基本大类展开的可视化检索界面。

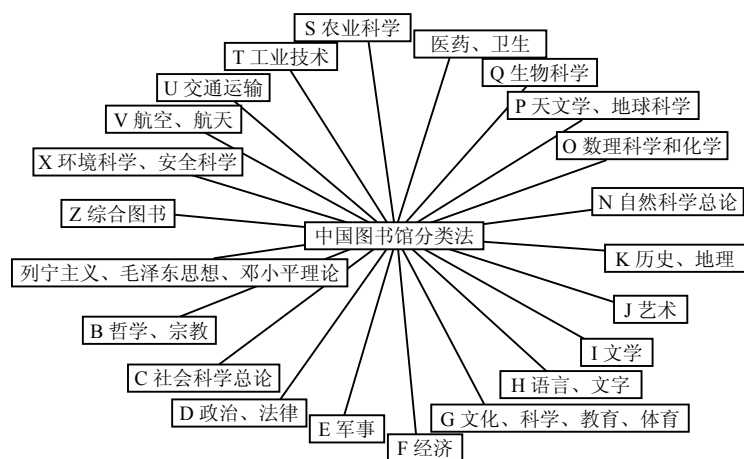


图 1.5 《中图法》可视化检索界面

(3) 按语义组织方式，信息组织可分为语义网技术、网络知识组织系统等方法与技术。

语义网并非是一个另外独立的 Web，而是现在的 Web 的一个延伸。在其中，所有的信息都有定义完好的含义，更利于人与机器之间的合作。语义网便于解决网络信息爆炸性增长所带来的巨大的处理和搜索问题。语义网的核心问题是如何有效地表达语义信息，使之能

够被计算机有效处理。传统的数字信息组织以数据层面的组织和信息层面的组织为主。在语义网环境下,数字信息的组织以知识组织为主。语义网技术和本体技术的结合使信息组织进入到更深的层次。语义网将改变传统互联网只是实现计算机硬件和网页的连接,而数据和信息资源分散在各网站的情况,对信息进行解释、交换和处理,更关注网络结构和语言的设计,使分布于全球的孤立的数据库融合,最终使用户独立运用 Internet 上庞大的信息资源。本体使知识组织体系从传统的树型结构向网状结构进化,为各类不同知识体系的结构和结合方式提供理论基础,很可能会极大促进信息/知识检索和导航功能的更新换代。

知识组织系统是各种对人类知识结构进行表达和有组织的阐述的语义工具的统称,包括传统图书馆建立在文献单元基础上的分类法、标题表、叙词表及更泛指的情报检索语言、标引语言,也包括网络时代建立在概念单元或知识单元基础上的知识地图、语义网、概念本体等。从以文献单元、数据单元为基础的知识组织系统发展到以本体为基础的语义网络知识组织系统,反映出知识组织系统研究的方法与技术不断提高、应用范围不断扩展。网络知识组织系统的出现由两个方面的共同作用所致:一是随着人类信息活动由纸制环境向数字环境迁移,传统知识组织系统的数字化、网络化势在必行;二是网络环境下信息量急剧增加,需要相应的语义工具实施对网络信息资源的组织,改进检索性能,实现对网络资源的深度挖掘和智能利用。网络知识组织系统代表了知识组织系统的发展趋势:数字化、网络化、语义化、协议化和自动化。网络知识组织系统的发展已经超越了“知识组织系统在网络环境中的应用”这一初级阶段,成为网络环境下新生的“语义工具”。网络知识组织系统将为图书馆知识组织的传统带来生机。从关键词查找发展到内容检索,其中网络知识组织系统的作用十分关键。

1.4.4 信息组织的未来发展趋势

互联网及信息技术的发展使信息传播方式发生了巨大的变化,对信息组织理论和技术则是一个重大挑战。面对庞杂无序多样化的数字信息,未来的信息组织应该以用户为中心,遵循实用性和易用性的原则,在继承优秀传统的同时,充分利用新的理论、方法、手段和技术,朝着更符合用户需要的方向发展。总的来看,应当朝以下方向发展。

1. 以需求为导向的信息组织方法和技术不断加强

1) 基于个性化服务的信息组织

随着信息技术的发展和用户需求的变化,个性化信息服务的趋势愈来愈强劲。这是目前网络信息组织努力发展的重点之一。信息组织个性化的努力可以包括多个层面的努力,包括专门搜索引擎的编制、综合性检索工具中特殊资源服务功能的建立、各种检索选择的提供(如大型搜索引擎中提供的对域名、语言、资源类型、检索网站的选择)、检索偏好设置的提供、检索优化的采用等,这都属于改进检索个性化的努力内容。但是,更深层次上的检索个性化,则利用网络搜索引擎的能力,发展检索个性化技术,通过对用户检索习惯和相关特点的分析,有针对性地提供相应的服务。

目前,基于网络的个性化信息组织与服务的思想在国外网站设计或信息服务系统中已经有一定的发展。雅虎推出了面向儿童的搜索引擎。许多门户网站和信息提供者推出了个性化定制服务系统,如 My Yahoo!、My Library!等。又如 Google 等使用的结合检索地点的资源提供技术等,在一定程度上反映了检索个性化的一种努力。国内许多网站为了满足国内读者和英语国家用户的不同需要,将其主页以多种版本组织(简体中文、繁体中文与英文版),这些做法体现了以用户为中心的思想。面向不同对象的信息组织也受到了图书馆界的重视,他们开发出了 My Gateway、My Library 等个性化定制服务系统,如美国北卡罗来纳州立大学图

书馆的 My Library，麻省理工学院和中国科学院都已建立起的基于用户的数字图书馆。国内部分高校图书馆也引进或开发了个性化定制服务系统。

在科学界，对信息组织的要求则更为专深、复杂，随着信息管理相关技术在各学科中的逐步应用，新知识发现可以通过信息管理相关技术得以实现，这使得科学家们对信息组织有了浓厚的兴趣和更高的期待。以行业领域数据挖掘为主的大数据和云计算成为新时期信息组织的重要研究领域，它们以处理海量多类型数据、高密度经济和研究价值等优势在学术研究、电子商务、数字图书馆、天文科学、大气科学、生命科学、农业科学、医疗卫生科学等多重领域广泛应用。

随着 Web 数据库技术、信息推送技术和智能代理技术等个性化信息服务所需的支撑技术的逐步成熟，网络信息组织将发展成一种信息代理服务，即根据网络信息的属性和用户需求，对网络信息进行加工、整理、排列和组合，使之有序化，以满足用户对网络信息的需求。

2) 信息检索工具的易用化

用户对各类信息、知识有日益增多的需求，但目前复杂的检索语言、词表结构和计算机技术却令许多人望而却步。以往的信息检索系统多是面向专业人员的，而对于普通用户来说，要适应计算机系统的键盘和编码方式，还需要经过一个再学习的过程。尤其是对于中国用户来说，以西方使用方式和英语为基础设计的计算机给他们的使用增加了困难。为使更多的人能享受到信息化带来的好处，使更多的普通人也能不费力地从检索系统中查到所需信息，要求提供更简便易行的信息组织和检索工具。目前已提出的解决信息组织工具易用化的方案有如下几个。

(1) 自然语言的使用。用户不必事先掌握复杂的词表结构和检索规则，用自然语言标引检索，做到“所想即所得”。开发出便于中国人识别和检索的中文信息组织和检索系统。

(2) 用户界面可视化。用用户易于理解的点线图、直方图、拼图、网状图及动画、三维技术和虚拟现实技术等表现复杂数据及其相互关系。例如，利用具有导航功能的信息地图，可将复杂、抽象的语义以直观的方式呈现给用户，与搜索引擎相比，这样不仅可以更清楚地揭示信息方位，而且能体现信息间的联系，帮助用户从全局了解信息的分布情况。

(3) 用文字识别、声音识别、触摸屏输入等更简单的人机界面代替目前的键盘输入等技术操作，清除用户学习的障碍。由于用户通过终端的屏幕获得网络信息资源，在进行信息组织时，还要考虑到对显示器屏幕的有效利用，以使用户能够更方便地利用网络信息。随着计算机智能化和超文本检索技术的完善，用于信息组织的智能软件会有很大发展。将来用户从信息组织到信息检索的过程有望像使用电视机、电话那样方便。

3) 信息内容揭示的深入化

数字时代用户对信息检索的需求不再仅仅满足题名、作者、主题词等传统条件下有限的检索点，而更注重实际内容（如目次、提要、文摘、全文、知识点等）的检索。这就对信息组织提出了更深入的要求。

(1) 组织的对象深入化。迅速普及的网络技术和数字技术使任意层次的任意信息元素、信息单元和信息集合体系正在逐步以计算机可识别和可理解的方式被定义、描述、指向、链接、传递和动态组织。网络信息组织的对象不仅停留在对信息特征的描述上，而且深入到知识单元，扩大标引广度，增加数据库的标引深度，通过多层次、多方位的描述与分析来揭示与组织网络信息资源，以促进网络信息资源的合理利用。

(2) 网络信息挖掘。目前网络上有 90% 的可用信息是非结构化信息，且网上数据往往是经常变动的和不规则的。网络信息挖掘在已知数据样本的基础上，通过归纳学习、机器学习、统计分析等方法得到数据对象间的内在特性，据此采用信息过滤技术在网络中提取用户感兴

趣的信息或更高层次的知识和规律。它除了处理传统数据库中的数值型的结构化数据外,处理更多的是文本、图形、图像、WWW 信息资源等半结构化、非结构化的数据。人工智能技术和信息推送技术促进了对网络信息的挖掘与深层次揭示,以更好地满足不同用户的各种需求。

(3) 向知识组织发展。未来的网络信息组织要更严格地控制信息的质量,对网上信息进行有效评价和筛选,为用户提供有价值的信息,而不是大量的无用信息,其目的是向人们提供便于利用的、可以帮助解决问题的、序化的知识,实现从信息层次到知识层次的根本转变,组织的知识包括显性知识与隐性知识。目前的知识组织主要以文献单元和以数据(各种事实、概念、数值的总和)单元为基础,但都是静态的、列举式的。未来的知识组织将以专家系统为基础,具有动态联系、判断、分析、比较、推理等新型知识处理与组织功能。

2. 信息组织工作的标准化与合作化

1) 信息组织标准化

在信息组织数字化和网络化过程中,必然遇到数据库、联机系统、检索系统和检索语言的兼容化和标准化的问题。例如,许多数据库的类型不同所引起的主题内容、学科领域、数据来源、文档格式、文档记录方式、标引规则及检索语言、输出格式等的不同给用户的使用带来了众多麻烦。影响了信息资源的共享。网络信息组织中的文件、搜索引擎、编目、学科信息门户等也都涉及标准化问题。因此,解决信息组织和检索系统的兼容和标准化问题成为信息组织发展的主要趋势。

(1) 图书情报机构的参与。国际图联制定了《FILA 书目记录的功能要求》。OCLC 一直注重研究、宣传与推行书目活动标准。国际知识组织协会(ISO)制定了知识与信息组织的相关标准,并在其主页发布。美国国会图书馆于 2002 年 5 月提出了“元数据输入与传输标准”,并设立了 Z39.50 维护机构和 MARC 办公室。美国图书馆协会的兴趣与活动之一是标准化与指导,并有专门网页提供标准化信息。我国在文化部的召集下,本着统一的规划、统一的技术标准及统一的运行规则等原则,组建了“中国数字图书馆工程建设联席会议”来协调工程的资源建设和标准规范。

(2) 国际和各国标准机构的推动。SC9 是国际标准化组织的 ISOTC46 的分委员会,它负责发展和维护关于文献展示、识别与描述的国际标准,并经常在其网站公布关于电子文献(包括网络信息资源)书目控制各方面的 ISO 国际标准草案。ISO 还成立了元数据工作组,负责元数据的标准与规范工作。美国全国标准化协会下设全国信息技术标准委员会,从事有关元数据的命名、标识、定义、分类和注册等工作,还成立了信息基础设施标准座谈小组。欧盟和英国的相关机构有信息社会标准化系统(ISSS)和英国标准协会的向用户传递信息解决方法部。另外,互联网管理机构、不同学科领域的学(协)会、公司、民间自发组织甚至个人,都在为网络信息组织及其相关的信息交换、信息检索、通信协议等方面标准的制定与推行作出积极努力。万维网联盟(W3C)是万维网上最有影响的互联网标准的认定机构,在网络信息组织领域,该机构认可的网络资源描述语言为 XML,资源描述框架为 RDF,元数据标准为都柏林核心元数据(DC),日期与时间格式为 W3CDTF。

国内外在网络信息组织标准化方面已取得一些进展,但仍有必要形成一系列标准与准则,使网络信息组织活动有规则可依。

2) 信息组织合作化

互联网是一个无主管的分散型互连结构,网络信息的通畅流动与有效利用要求各方面的整体配合。网络信息的组织是一项涉及面很广的持久性工作,需要世界范围内的合作以保证准确、及时地报道网上信息及其变动情况,提高信息的质量并实现规模效益。网络信息组织

的合作将得到强化。

图书情报界已经开展了网络信息组织合作化的有关活动。国际图联发起了“全球书目控制和国际机读目录核心活动”，并编辑出版了《国际编目与书目控制》杂志。2002年7月在西班牙召开的第7届国际知识组织大会的主题是“21世纪知识表示与组织的挑战：跨越边界的知识一体化”。OCLC一直在书目控制与资源共享的协作中扮演着国际中心的角色，国际图书馆协作体联盟则是将各种协作体组织起来的组织机构。又如，Renardus是欧洲范围内开展的合作项目，该项目由欧洲用户友好信息社会项目发展而来，得到了欧洲各国家图书馆、大学研究和技术中心、全欧洲主题信息网关的支持和合作。自2000年1月以来，这些合作者们正努力构建一个单一的基于Web的“中介服务”，该服务的目标在于为遍布欧洲的分布式网络科学和文化信息资源提供跨库检索和跨库浏览。

网络信息组织的合作还将超出图书情报界，扩大到整个信息生产链上的其他所有参与者，包括网站内容创作者、出版商和信息资源系统的合作和集成者等。美国的网络图书馆是世界上最大的全文电子图书的收藏者与服务提供者，也是出版商、发行商、图书馆与读者有效结合的典范。《出版机构与国家书目服务连接》是由英、荷、挪、法、西班牙5个国家图书馆等9个机构合作的项目。该项目解决了出版机构书目数据与MARC数据的双向转换，有利于实现将出版机构提供的数据作为电子出版物网络元数据的构想。

3. 信息系统的互操作和信息组织大众化

1) 信息检索系统的互操作

在分布、异构、变化的网络环境下，数字信息资源数量巨大且高度分散，而信息源、数据格式、用户需求高度异质，对如此庞杂的信息进行格式化和结构化很困难，用户要找到真正符合需要的信息也很困难。如何将广泛分布的、自治的、异构的信息资源和信息检索系统联合起来，向用户提供统一、透明的服务并实现信息系统的互操作，成为当前和未来信息组织研究实践的热点，引起了普遍关注。

系统互操作是指分布信息检索系统之间能无缝地交换、共享信息资源和信息服务，并能在不损害各个分布系统自主性的同时构成一个虚拟的集成系统。在信息资源组织与服务的集成中，信息系统互操作的目标是向用户屏蔽分布的、异构的各个信息系统间的差别，实现用户对多个信息系统的交叉浏览和交叉检索，提供统一入口的多个信息系统的检索和浏览服务，实现信息共享。

信息资源组织与集成服务要解决互操作的以下主要问题：屏蔽分布的各信息系统之间的差别，为用户提供一个一致的服务，在统一界面上进行的跨仓储的服务对于用户来说是透明的；为信息资源和信息系统提供一种灵活的集成机制，这种集成机制必须允许各个相对独立的信息系统自由增加新的服务，或对以前的服务进行修改；信息资源整合和集成服务协议制定，包括元数据协议、数字对象存储协议、信息搜索协议、付费协议、信息资源服务的运行管理协议等；开发信息资源整合与集成服务系统高层协议中间件，实现分布子系统间各项服务的互操作。

2) 信息组织大众化

Web2.0及相关概念的引入改变了现有万维网信息提供的模式，使信息组织不再局限于专业人员的范畴，而是走向大众化。

目前网络上的信息检索和传输系统大都采用服务器和客户器的关系模式，服务器总是被动地等待客户的信息需求做出相应的响应。在Web2.0中，用户已经不再简单地只是信息的获取者，同时也是信息的提供者。Web2.0发展了一系列新的形式，包括博客、互动标签(Tag)、社会网络交往系统(SNS)和资源订购系统(RSS)等。这类新形式除了使信息大量增加外，

更加重视用户作用的发挥,极大地促进了信息的传输和交流,而且在信息资源表述和关系揭示上具有重要作用,表现在以下几个方面。

其一,通过用户的参与,改进资源揭示。最典型的是 Tag 中用户个性化标注的使用,这一形式使用户的检索表述得以反映。

其二,利用用户交流和使用揭示信息资源之间的联系。SNS、RSS 都具有这一作用。这些联系是用户根据需要自动建立的,融入了用户的智力判断。

其三,通过交互方式的应用实现整合,例如,Tag 中某个特定用户个性化标注的使用有可能是不适用的,但通过在使用过程中用户的调整和整合,适用的表述会得到保留;RSS、SNS 中联系的建立等也具有同样的情况。尽管个体的处理会出现错误,但作为整体、作为统计学意义上的多数的判断,必然有其合理性。

Web 2.0 为资源处理引入了新的因素,改变了资源的数据特点,这些变化的实质是,用户的参与和用户智力判断的加大,而这一变化是结合使用需要进行的,是以最为自然的、人为性最少的方式实现的。Web 2.0 的出现将会对网络资源的组织方法、处理技术等产生深刻的影响,包括资源搜索软件中采集策略的设置、检索提供中排序因素的确定、相关性揭示中相关性权值的设置,并有可能衍生出新的检索系统,如基于互动标签的浏览系统、基于 RSS 或 SNS 的相关揭示系统等。因此,Web 2.0 的发展可能将对上述各个方面的发展都带来不同程度的影响和促进。

Web 2.0 提供了信息自组织功能,其兴起和应用创造了一个全新的信息空间。Web 2.0 将复杂的技术移至后台,在简单规则的约束下,用户广泛参与。知识信息的生产、传播和利用在多元化、多样化、个性化和去中心化模式下实现了自组织和有序化。从博客信息交流社区的形成到维基百科的协同组织编辑及社会化书签产生的分众分类等,无不体现了 Web 2.0 的信息自组织功能和序化特征。

4. 信息组织的智能化和语义网格化

1) 信息组织智能化

网络环境下,先进的信息技术将部分地代替人脑进行信息组织中的信息识别、信息分析综合和信息重组,进而实现智能程度更高的“知识组织”,即用高度模拟人脑思维机制与习惯的方法来组织知识。

人工智能技术将促进网络信息的深层次挖掘和揭示,更好地满足用户的不同需求,如系统自动运行、不断更新用户资料库、提供个性化的主动服务等。计算机将能理解人们用自然语言输入的信息并正确回答用户提出的有关问题,还能自动对输入的信息做摘要,能用不同词语复述输入的内容,还能把用某一种自然语言表示的信息自动翻译为另一种自然语言。其中,使用最广泛的智能化技术是超文本、专家系统、数据仓库、知识挖掘等。

目前,有关信息组织智能化的一些关键技术的研究如自然语言理解、知识的表示、知识的获取等已取得一定成果,出现了自动抽取、收集和产生元数据的软件。例如,利用智能化的书签软件可以从 URL 中解析出元数据,并自动进行分类标引,还可通过监测用户的信息搜索与浏览过程,自动获得用户的需求与兴趣信息,把合适的信息提交给用户,并允许用户订购感兴趣的新资源或更新已有的信息资源。随着网上自动分类、自动标引、自动编制分类表与词表、自动漫游技术、信息类别的自动判别技术和信息推拉技术的逐步发展完善,会有越来越多的网络信息资源被自动地追加、组织到相应的位置,方便用户及时准确地检索。同时,信息组织技术的智能化将使信息组织工作变得十分简便,它不再仅仅是专家和技术人员的事情,更多的用户将成为信息组织活动的直接参与者。

2) 信息组织语义网格化

近年来,网络结构管理中的一个新进展是语义网格(Semantic Grid)技术的发展。网络计算(Grid Computation)及其相关的语义网格(Semantic Grid)技术是着重从计算机网络结构的角度的来研究和开发计算资源及进行语义处理的。网络计算最初是针对复杂的科学计算伴随着互联网的发展提出来的,被看成是一种新型计算模式。这种计算模式利用计算机网络把分散在不同地理位置的计算机组织成一个“虚拟的超级计算机”。每一台参与计算的计算机被看成是一个“节点”。这样,整个计算就是由成千上万个“节点”组成的“一张网格”,故这种计算方式被称为网格计算。网格计算在较长的一段时间里都是独立于语义网而发展的。现在语义网格技术也采用了语义网的许多技术标准,如RDF/PDPS语言标准,用于对应的网络资源描述。语义网格技术正在呈现出与语义网技术相结合的趋势,为网络资源的自动计算及其组织体系提供了一个值得关注的方向。

随着语义网格概念的提出,国内外一些学者试图将这种新技术应用于数字信息的组织研究中。英国Ali Shiri在E-Science项目报告中提出利用本体为数字资源构建基于网格的语义框架。斯洛维尼亚Liubljana大学的Ziga Turk指出,语义网格技术对于识别和标注概念和术语、本体描述、体系结构构建及用户需求分析起着至关重要的作用,并指出语义网格技术在互操作、数字资源和虚拟组织等方面扮演着重要的角色。

总之,未来的信息组织发展的趋势将会呈现跨学科、跨国界、跨地域和不同文明之间高度融合,技术应用高度智能化,在不同文化背景中深度兼容化与走向标准化,理论研究高度深入化,信息服务个性化和泛客户化,信息资源和背景高度复杂化,所涉及内容向纵深化方向发展。



本章小结

信息是社会发展的一个重要前提和主要资源,是人类的宝贵财富。信息资源是现代社会最重要的战略资源之一,它的开发利用水平是国家综合国力的体现。信息组织就是根据人类社会发展的需求,以各类信息源为对象,通过对其内容特征等的分析、选择、标引、处理,使其成为有序化集合的信息增值活动,是信息资源开发和利用的重要环节。



问题讨论

1. 信息组织的含义和特点是什么?
2. 信息组织包括哪些内容?
3. 按信息的运动状态和方式可以把信息组织划分为哪几个层次?
4. 中外信息组织发展历史有哪些异同?
5. 信息组织研究前沿领域有哪些?




第2章

信息组织的理论与方法基础

本章引言

信息组织作为一种社会实践活动，是建立在一定的理论基础和方法基础之上的。人们在长期的信息组织活动中，不断从相关学科的理论中汲取营养，典型的如有序化理论、信息构建理论、知识论、本体论、分形理论等，这些理论为信息组织活动奠定了理论基础。同时，语言学、逻辑学和知识分类学则为信息组织提供了方法基础，无论是信息组织的理论基础还是方法基础，它们都对信息组织活动具有不可或缺的指导作用。

本章重点

- 有序化理论；
 - 信息构建理论；
 - 知识论、本体论和分形理论；
 - 信息组织的语言学基础；
 - 信息组织的逻辑学基础；
 - 信息组织的知识分类学基础。
- 

2.1 信息组织的理论基础

信息组织是将无序的特定信息，根据一定的原则和方法，使其成为有序状态的过程，以便有效地传递和利用信息。人们在长期的信息组织实践中，不断地从相关学科的理论，尤其从近几十年来兴起的有序化理论（包括自组织理论即新三论）、信息构建理论、知识论、本体论、分形理论等中汲取营养，为信息组织活动奠定了理论基础。

2.1.1 有序化理论

序是事物的一种结构形式，是指事物或系统的各个结构要素之间的相互关系及这种关系在时间和空间中的表现。所谓有序，是指事物内部的要素或事物之间有规则的联系和运动转化，当事物结构要素具有某种约束性且在时间序列和空间序列呈现某种规律性时，这一事物就处于有序状态；反之，则处于无序状态，即事物内部各种要素或事物之间混乱而无规则地组合和运动变化。有序与无序在一定的条件下统一形成事物的秩序。古人早就对世界的秩序进行过直观的猜测。中国古代和古希腊的思想家们都认为，世界是从毫无秩序的一片混沌发展来的。近代以后，在自然科学研究的基础之上，许多学者，如狄德罗、康德等也都持世界从无序走向有序的观点。在现代，学者们更加广泛地探索有序和无序的关系问题。目前，对于有序与无序的转化的研究，已成了现代科学中一个难点和热点问题，探讨这些问题有助于科学的进步和人类认识的深化。

一般认为有序化理论包括老三论和新三论。老三论是指系统论（System Theory）、控制论（Cybernetics）和信息论（Information Theory），是20世纪40年代先后创立并获得迅猛发展的三门系统理论的分支学科。虽然它们出现仅有半个世纪，但在系统科学领域中相对于“新三论”已是元老。人们摘取了这三论的英文名字的第一个字母，把它们称之为SCI论。新三论是指耗散结构论（Dissipative Structure Theory）、协同论（Synergetics）、突变论（Catastrophe Theory），是20世纪70年代以来陆续确立并获得极快进展的三门系统理论的分支学科，也称为DSC论，又被称为自组织理论。无论是“老三论”还是“新三论”，它们都是进行信息组织的理论基础来源，其中与信息组织关系最为密切的则是系统论、“熵”与“负熵”及信息论、自组织理论中的耗散结构论和协同论等理论。

1. 系统论

系统一词来源于古希腊语，指由部分构成整体的意思。系统论是研究系统的一般模式、结构和规律的一门学问。它研究各种系统的共同特征，用数学方法定量地描述其功能，寻求并确立适用于一切系统的原理、原则和数学模型，是具有逻辑和数学性质的一门新兴的科学。作为一门科学的系统论，人们公认是由美籍奥地利人、理论生物学家贝塔朗菲（L.Von.Bertalanffy）所创立的，他在1952年发表“抗体系统论”，提出了系统论的思想。1973年提出了一般系统论原理，奠定了这门科学的理论基础。

系统论认为，整体性、关联性、等级结构性、动态平衡性、时序性等是所有系统共同的基本特征，也是系统所具备的基本思想和系统方法的基本原则。系统论的核心思想是系统的整体观念，任何一个系统都是一个有机整体，而不是各个组成要素的简单相加或机械组合，同时，系统中各要素不是孤立存在的，每个要素都处于特定的位置，发挥着一定的作用。系统是由相互作用和相互依赖的若干组成部分结合成的、具有特定功能的有机整体，而系统本身又是它所从属的一个更大系统的组成部分。复杂事物的功能远大于某组成因果链中各环节

的简单总和,认为一切生命都处于积极运动状态,有机体作为一个系统能够保持动态稳定是系统向环境充分开放及获得物质、信息、能量交换的结果。系统论强调整体与局部、局部与局部、系统本身与外部环境之间相互依存、相互影响和制约的关系,具有目的性、动态性、有序性三大基本特征。系统论的基本思想方法,就是把所研究和处理的对象当做一个系统,分析系统的结构和功能,研究系统、要素、环境三者的相互关系和变动的规律性。

当今社会已经步入信息社会,信息成为世界三大资源之一。著名未来学家奈斯比特曾说:“失去控制和无组织的信息在信息社会里不再构成资源。”信息组织的主要研究内容之一就是如何将信息社会中的大量分散、无序的信息组织成一个有序的系统,并在它们之间建立起内在的相互联系。从系统论的观点出发就是要使有组织的信息整体的功能大于各个信息单位的功能之和,这样信息的效用才能够得到充分的发挥和使用。

系统论反映了现代科学发展的趋势,它的出现使人们的思维方式发生了深刻的变化。系统论不仅为现代科学的发展提供了理论和方法,而且也解决现代社会中的政治、经济、军事、科学、文化等方面的各种复杂问题提供了理论和方法基础,作为现代科学的新潮流,系统观念正渗透到每个领域,促进着各门科学的发展。

2. “熵”与“负熵”及信息论

“熵”(Entropy),指的是混乱的程度,是由德国物理学家克劳修斯(Rudolf Clausius, 1822—1888)在1854年创造的一个术语,熵最早是热力学上的一个符号,是一种测量在动力学方面不能做功的能量综述,表达的是某一系统内部热量平均化的程度,亦被用于计算一个系统中的失序现象。“熵”在经典热力学中通常用符号 S 表示,可用增量定义为 $dS = (dQ/T)$,式中 T 为物质的热力学温度; dQ 为熵增过程中加入物质的热量。

“熵”这个概念以后被许多其他学科所借用,引申出更多的概念。1865年,玻耳兹曼(Boltzmann)从统计力学的角度定义了“熵”: $S = k \ln \Omega$,式中 k 是玻耳兹曼常数, Ω 是系统某一宏观状态所对应的微观状态的数目即热力学概率,它的微观意义是系统内分子热运动无序性或混乱程度的一种度量。1948年克劳德·艾尔伍德·香农(Shannon)第一次将“熵”引入到信息论中来,在信息论中,“熵”可用做某事件不确定度的量度。信息量越大,体系结构越规则,功能越完善,熵就越小。利用熵的概念,可以从理论上研究信息的计量、传递、变换、存储等活动。

熵在信息论中的定义如下:如果有一个系统 S 内存在多个事件 $S = \{E_1, \dots, E_n\}$,每个事件的概率分布 $P = \{p_1, \dots, p_n\}$,则每个事件本身的信息为: $I_e = -\log_2 p_i$ (对数以2为底,单位是位元(bit))、 $I_e = -\ln p_i$ (对数以e为底,单位是纳特(nat))。例如,英语有26个字母,假如每个字母在文章中出现次数平均的话,每个字母的信息量为

$$I_e = -\log_2 \frac{1}{26} = 4.7$$

而常用的汉字有2500个,假如每个汉字在文章中出现次数平均的话,每个汉字的信息量为

$$I_e = -\log_2 \frac{1}{2500} = 11.3$$

整个系统的平均信息量为

$$H_s = \sum_{i=1}^n P_i I_e = -\sum_{i=1}^n P_i \log_2 P_i$$

这个平均信息量就是信息熵。

负熵是物质系统有序化、组织化、复杂化状态的一种量度。1944年，著名的物理学家、量子力学的奠基人之一、诺贝尔奖获得者薛定谔（E.Schrodinger）出版《生命是什么？》一书，明确地论述了负熵的概念。他认为，“既然 Ω 是无序性的量度，那么它的倒数 $1/\Omega$ 可以作为有序性的一个直接量度，因为 $1/\Omega$ 的对数正好是 Ω 的对数的负值，玻耳兹曼方程可写成 $-S=k\ln(1/\Omega)$ 。因此负熵的表达式可以换成一种更好一些的说法：取负号的熵，它本身是有序的一个量度。

在信息论中，学术界目前普遍接受的观点是“信息即负熵”，有少量学者提出相反的观点：“信息即熵，非负熵”。从信息的基本定义出发，或者从信息的作用出发，信息是用来消除人们对事物的存在方式和运动状态的不确定性的，那么“信息即负熵”似乎更有道理、更具有实际意义；关于信息与“熵”、“负熵”之间的关系，随着不同观点和看法的提出，还有待于我们对信息科学的进一步研究和论证。

3. 自组织理论

自组织理论又称为“新三论”，是20世纪70年代后开始建立并发展起来的一种系统理论。它的研究对象主要是复杂自组织系统（如生命系统、社会系统）的形成和发展机制问题，即在一定条件下，非线性的系统是如何自动地由无序走向有序，由低级有序走向高级有序的。它主要由三个部分组成：耗散结构理论、协同论和突变论。

德国理论物理学家H. Haken认为，从进化形式来看，组织可以划分为两类：他组织和自组织。如果一个系统靠外部指令形成组织，就是他组织；如果不存在外部指令，系统按照相互默契的某种规则，各尽其责而又协调地自动地形成有序结构，就是自组织。自组织现象无论在自然界还是在人类社会都是普遍存在的，一个系统的自组织功能愈强，其保持和产生新功能的能力也就愈强。近年来，随着信息数量指数级的增长和信息技术的飞速发展，信息系统已经开始逐渐具备自组织的条件，特别是一些网络信息，已经具备自组织的开放性、远离平衡和非线性等相关特征，因此，将自组织理论作为信息组织的理论基础是十分必要的。

1) 耗散结构理论

20世纪60年代末，著名比利时物理学家普利高津（I. Prigogine）提出了一种关于非平衡系统自组织的理论——耗散结构理论。耗散结构的概念是相对于平衡结构的概念提出来的。长期以来，人们只研究平衡系统的有序稳定结构，并认为倘若系统原先是处于一种混乱无序的非平衡状态，是不能在非平衡状态下呈现出一种稳定有序结构的。

耗散结构理论指出：一个远离平衡态的非线性的开放系统（不管是物理的、化学的、生物的乃至社会的、经济的系统）通过不断地与外界交换物质和能量，在系统内部某个参量的变化达到一定的阈值时，通过涨落，系统可能发生突变即非平衡相变（质变），由原来的混沌无序状态转变为一种在时间上、空间上或功能上的有序状态。这种在远离平衡的非线性区形成的新的稳定的宏观有序结构，由于需要不断与外界交换物质或能量才能维持，因此称之为“耗散结构”。

一个典型的耗散结构的形成与维持至少需要具备三个基本条件：首先，系统必须是开放系统，孤立系统和封闭系统都不可能产生耗散结构；其次，系统必须处于远离平衡的非线性区，在平衡区或平衡区都不可能从一种有序走向另一更高级的有序；第三，系统中必须有某些非线性动力学过程，如正负反馈机制等，正是这种非线性相互作用使得系统内各要素之间产生协同动作和相干效应，从而使得系统从杂乱无章变得井然有序。信息组织通过与外界交换物质、能量与信息，对信息加工整序使信息系统成为远离平衡态的开放系统，并具有输入、输出、多次循环及反馈等基本特征，因此耗散结构理论可以作为信息组织的

理论基础之一。

2) 协同论

协同论, 亦称协同学或协和学, 由联邦德国斯图加特大学教授、著名物理学家哈肯(Hermann Haken)创立。1971年他提出协同的概念, 1976年系统地论述了协同理论, 发表了《协同学导论》、《高等协同学》等著作。

协同论主要研究远离平衡态的开放系统在与外界有物质或能量交换的情况下, 如何通过自己内部协同作用, 自发地出现时间、空间和功能上的有序结构。协同论以现代科学的最新成果——系统论、信息论、控制论、突变论等为基础, 吸取了结构耗散理论的大量营养, 采用统计学和动力学相结合的方法, 通过对不同领域的分析, 提出了多维相空间理论, 建立了一整套的数学模型和处理方案, 在微观到宏观的过渡上, 描述了各种系统和现象中从无序到有序转变的共同规律。

协同论是一门研究系统进化普遍规律的科学, 它研究由许多子系统构成的系统如何通过协作从无序到有序演化的规律。信息组织的目标就是使信息单元由无序集合转换成有序集合, 如何建立各个信息单元之间的协同作用机制关系到信息组织的成效, 因此, 协同论是信息组织的理论基础之一。

2.1.2 信息构建理论

1. 信息构建的含义(包括起源与发展)

信息构建(IA, Information Architecture)最早由美国建筑师沃尔曼(R.S. Wurman)于1976年提出, 他将信息(Information)和建筑学(Architecture)相结合, 创造了“信息构建”这一词汇, 他认为:“在满足使用者需求这一点上, 构筑信息建筑物与构筑物理建筑物有相同之处, 它们都可以看成是一种服务于特定目标的建筑设计工作。”但当时由于各种信息条件和环境的制约, 他关于信息构建的思想和认识并没有引起社会的广泛关注。直到20世纪90年代中后期, 随着计算机技术和网络技术的迅速发展和普及, IA逐渐引起人们的广泛关注, 具体表现为: IA知识成为网站设计和开发的必备知识, 有关IA的学术专著、论文和报道大量出现, 有关IA的协会和组织也纷纷成立, 以培养信息建筑师为目标的高等教育和职业培训开始出现等。

目前, 关于信息构建的含义学术界尚未达成共识。沃尔曼先生认为:“IA是建设信息结构, 让其他人理解。”他用建筑学的隐喻来描述自己作为信息建筑师的工作:“我将信息建筑师看做是生成系统的、结构化的、有序的原则的人, 他们让事物工作起来, 让人造物品、思想和政策因其具有清晰性而能起到告知的作用。”信息构建研究会(AIFIA, Asilomar Institute for Information Architecture)将信息构建定义为:“是共享的信息环境结构的设计; 是组织和标识网站、内联网、联机交流和软件以保证其可用性和可找到性的艺术和科学; 是一个致力于对数字园区(Digital Landscape)的设计和建设的、正在出现的实践领域。”美国信息科学技术学会在2000年IA峰会上提出的信息构建定义式:“信息构建是组织信息、帮助人们有效地实现其信息需求的艺术和科学。”

信息构建理论旨在满足用户信息需求的新的理论构想, 它从一个新的角度, 重新构建和整合了有关信息组织的有关知识, 较通俗地解释了信息爆炸情况下信息的收集、标引、分类和利用等问题, 并对网站门户、各种数据库的信息组织和利用的实践具有较强的指导作用。IA的五规则就对数据库的精选信息和精选期刊有指导意义。例如, 规则4: 确信哪些信息值得保留, 你真正想要了解哪些信息; 规则5: 大多数信息是无用的, 要勇于放弃无用信息。

国内学者将信息构建概括为：“组织信息和设计信息环境、信息空间或信息体系结构，以满足需求者的信息需求，实现他们的目标的一门艺术和科学。”IA 的定义所包含的内在和外在的含义，对于图书馆学、情报学的研究而言，其突出意义在于 IA 给图书馆学、情报学研究提出了新的要求，而且这种要求完全是以一种新的角度提出的，即：

- IA 要求获得和掌握信息内容；
- IA 要求将信息组织好；
- IA 要求优化信息结构设计；
- IA 要求为读者表达信息内容；
- IA 要求提供一个清晰的易于获得信息的界面。

有的学者分析了信息构建过程中信息状态的变化特征，从而概括和归纳了信息构建理论与实践所体现的四条基本原理。

信息片段：信息片段是信息构建理论中一个重要的概念。信息由组成信息的基本元素，如文字、符号、图形、视频、音频等形式构成了具有某种含义的信息片段。信息片段由信息元素组成，是已经有基本含义的信息模块，是构成信息集合的组成要素。

信息集合：各种信息自然地或人为地聚集在一起，聚集在某种信息容器中，就构成了信息集合。信息集合是信息构建的基本对象。

信息结构：信息结构可以看成是对信息集合中的信息按照某种方式加以组织后所形成的有序的信息整体，它是信息系统存储信息和信息系统内外部交流信息的方式和渠道。

信息空间：信息空间是信息存在或发生的地点，可以指某个相关领域的信息的总和。信息空间既可以指物理存在的信息空间，又可以指作为人们认识状态存在的信息空间。

在信息构建活动中，作为信息构建主体的信息构建师（Information Architects）需要在对信息内容、信息用户的信息需求及信息环境理解的基础上，对随机地、自然地或人为地聚集起来的信息片段生成信息集合；在对信息集合中信息片段之间的各种关系进行分析的基础上建立信息结构；还需要进行信息界面设计，将信息结构及信息结构中包含的信息内容以科学的、艺术的方式，在特定的信息空间中将信息展示给用户。

由于信息爆炸与信息焦虑、信息饥渴，才导致了信息构建理论的出现。信息构建的主要活动是组织信息内容、生成信息结构和设计信息界面，其直接目标是建立一个清晰的、易于理解的信息结构，信息构建的最终要求是提供给信息的用户一个良好的信息空间环境。根据信息构建活动中信息状态的特点，可以总结出信息构建的四条基本原理。

（1）信息片段集成原理。信息片段集成原理指信息构建过程是从信息片段的采集开始，对采集的信息进行内容和谐的、各种媒介和手段兼容的、综合的、多方面的集成的过程。集成过程最重要的问题有三个：一是和谐地集成信息资源；二是有效使用不同的交流媒介或工具并使它们能够彼此兼容；三是实现小屏幕大集成的功能。

（2）信息集合序化原理。信息集合序化原理是指信息构建过程中对信息集合中信息内容的组织和信息形式的表达实质上是增强有效信息含量，自觉控制信息结构体系中的熵值，形成有条理、合逻辑、主题鲜明、主次关系清晰的信息结构体系。

（3）信息结构展示原理。信息结构展示原理指信息建筑师为序化后的信息设计一个协调一致的、功能化的信息架构，实质上是通过信息界面表达它们，有效地展示自己的内容、风格和特色，让用户能感知信息结构中存在的信息，可以方便地、心情愉悦地从中获得信息，满足自己的信息需求和实现自己的目标。

（4）信息空间优化原理。信息空间优化原理指信息构建过程中通过一系列手段和措施，在复杂庞大的信息空间中帮助人们缓解信息环境造成的心理上的迷惑或行动上的困境，减轻

人们认知负担, 加强人们信息感知和信息捕捉能力, 促进信息接收和利用。信息空间优化原理有宏观和微观两方面的表现形式^①。

2. 信息构建的应用

关于信息构建的应用, 由于 IA 繁荣、发展于网络环境, 所以很多学者认为 IA 主要是针对网站而言的, 也有学者认为 IA 无处不在, 即“哪里有信息, 哪里就有 IA”。IA 起源于任意的信息集合, 但受到网络环境的促进, 发展于网络信息的构建问题, 尽管它目前的主要应用领域是网络信息构建, 但它从来就不是仅仅针对网络的, 也不可能局限于网络信息的构建, IA 是针对所有信息集合的。

沃尔曼先生多次在著作中提到除了网络之外的不同的信息内容的信息设计和构建问题, 他对城市指南、烹调书内容设计、产品说明书、地图等多种不同的内容进行过内容结构的构建尝试; 谢尔·凯门 (Shel Kimen) 在网上发表的《关于信息构建的 10 个问题》中也指出: 从最基本的要素看, 信息构建就是建立一个结构或组织信息。在图书馆中, 信息构建指编目系统和保存书本的建筑物的物理设计的结合体; 在 Web 上, 信息构建指将网站内容组织成一些范畴, 并生成一个界面来支持这些范畴的结合体。

国内学者将 IA 的应用领域概括为以下四个方面:

- (1) IA 是针对所有信息结合而言的;
- (2) IA 出现在所有需要对信息集合中信息内容进行展示的信息空间中;
- (3) 尽管不同的信息集合形式对 IA 的依赖程度不同, 但 IA 的原理和方法都可以作为其信息组织和内容管理的指导思想和实用工具;
- (4) 在一般情况下, 陌生的信息环境和复杂的信息空间更需要 IA。

目前, 除了应用于网站开发, 人们已经将 IA 应用于企业、政府、学校等多个领域和部门, IA 已深入到我们日常生活中常见的信息集合中, 如图书、期刊、工具书、地图、图书馆、数据库、信息系统等。

3. 信息构建与信息组织

信息构建是关于如何组织信息, 以便帮助人们有效实现信息需求的艺术和科学, 即 IA 最根本的对象是信息。信息组织则是按照一定的目的、任务和形式对信息加以序化的过程。

虽然我国关于信息组织的研究早于关于信息构建的研究, 但是从研究对象、研究目的和本质上看, 信息构建和信息组织之间具备一定的一致性。诚然, 二者在其产生的社会背景、所采用的技术方法等方面也存在一些差别。

从研究对象上看, 信息构建和信息组织都以信息为研究对象, 信息是自然界、人类社会和人类思维活动中普遍存在的一切物质和事物的属性。与信息普遍存在相对应的现象即是“信息构建无处不在”。其次, 信息构建和信息组织的目的都是满足人们的信息需求。信息组织通过建立信息资源存储系统和检索工具, 使用户能够方便、有效地获取所需的各种信息资源, 它的最终目标是使信息有序化, 从而方便人们的利用、满足人们的信息需求; 信息构建则通过建立一整套的信息系统, 使信息清晰化和易于人们理解来满足人们的信息需求。所以, 从本质上来说, 信息构建和信息组织都是为了满足用户的信息需求而对信息进行处理的方式方法。

从产生的社会背景上看, 在我国, 最初的信息组织思想起源于图书文献整理的实践, 伴随着信息社会的到来, 信息数量呈现出指数级的增长, 而信息组织正是伴随着社会信息数量

^① 周晓英. 论信息集合的信息构建 (IA). 情报学报, 2004, 23(04): 456~462.

急剧增加、流速加快、分布散乱、优劣混杂等与使用矛盾现象日趋严重而发展起来的；进入20世纪90年代，互联网飞速发展，此时的信息环境与20世纪70年代大不相同，一方面，信息环境恶化；另一方面，表现在信息焦虑现象普遍存在。正是在此背景之下，信息构建理论应运而生，它通过组织信息、创建信息结构或地图，从大量处于复杂状态的信息中抽取本质模式，帮助人们理解信息，找到知识路径，有效解决了由于信息环境恶化而带来的一系列不良影响。

从采用的技术方法上来看，长期的信息组织实践已经创造出了多种多样的技术方法。目前已经形成了以分类法、主题法和元数据法在内的完整的信息组织的技术方法体系。而有关信息构建的研究开始得较晚，其有关技术、方法的研究也还在发展之中，但我们应该明白，信息构建是基于技术的，没有信息技术的发展就谈不上信息构建。可以说信息组织是从微观上来对信息进行处理的，而信息构建则更多地从宏观上对信息与信息内容及其结构进行研究。

2.1.3 知识论

1. 传统哲学研究中的知识论

“知识论”（Epistemology）来源于希腊语“知识（Episteme）”和“词/演讲（Logos）”，它是探讨知识的本质、起源和范围的一个哲学分支。在西方哲学中，古代与近代的哲学家有关知识论的研究主要是从人的认识能力的角度进行的，把有关认识的研究建立在人的感性和理性的基础上，这种认识理论是发生学意义上的，正是由于这种认识理论的发生学性质，致使国内哲学界有些学者将“Epistemology”译为“认识论”。知识论作为哲学的一个主要分支，其理论是与哲学的理论形态密切结合的。

2. 蒯因（又译奎因）的知识整体论（也称整体知识论）

蒯因的知识整体论的提出具有十分重大的意义，他通过“语义上溯”，在“观察句”和“观察陈述”层面上建立外部刺激与科学理论之间的证据联系来解决知识论问题。“语义上溯”是西方哲学发生“语言转向”后哲学研究的一个基本策略，它把原有认识论中关于外部对象的实在性及其真理性的哲学问题归结为对语言、对我们所说的句子的意义的理解，这样，我们关于外部世界的探讨就转化为关于语言的探讨了。

蒯因的整体知识论可以概括两个方面：“①我们不可能举出任何一个可避免经验反驳的句子（拒斥先验认识）；②在理论和预言矛盾的情况下我们绝不能指出某些引起这些矛盾的句子（拒斥孤立主义的单个假说检验）；相反，始终是作为整体的系统要么受到怀疑，要么又被调整好。”蒯因整体知识论的第一个结论是取消传统逻辑中综合命题和分析命题的区别；第二个结论是认为不存在传统的知识论，而只存在科学知识系统和对它的解释。蒯因整体的知识论是从迪昂的“不可能对一个孤立的假说进行检验”的思想发展而来的，但是真正导致蒯因整体主义知识论产生并走向自然化的知识论的是纽拉特的整体主义知识观。因此，我们一般把知识整体论称为“迪昂—纽拉特—蒯因”论题。

3. 钟义信的知识论：全信息及信息—知识—智能的统一理论

我国学者钟义信在定义一组关于知识理论的基本概念，即信息、知识和智能的基础上，揭示了这三者之间的相互联系，提出了信息—知识—智能的统一理论。他认为知识论的核心问题是揭示知识与信息、知识与智能之间的关系，并阐明“如何把信息提炼成知识；如何把知识激活成智能。”

可以从本体论层次和认识论层次来定义信息。所谓本体论信息，是指事物运动的状态及

其变化方式的自我表述。这里的“事物”，是泛指一切可能的研究对象，包括外部世界的物质客体，也包括主观世界的精神现象。“运动”泛指一切意义上的变化，包括机械运动、物理运动、化学运动、生物运动、思维运动和社会运动等。“运动状态”是指事物运动在空间上所展示的性态和形态。“运动状态的变化方式”指事物运动状态随时间而变化的过程样式。所谓认识论信息，是指主体所感知（或所表述）的关于该事物的运动状态及其变化方式，包括这种状态/方式的形式、含义和效用。信息是物质的一种属性，它不同于消息，也不同于信号、数据、情报和知识。信号是信息的载体，数据是记录信息的一种形式，情报通常是秘密的、专门的、新颖的信息，知识是认识主体所表述的有序化的信息。信息的外在形式、内在含义和价值效用三个因素应有机的进行统一处理，否则就不可能理解信息的本质。香农的贡献在于用概率熵（负熵原理）描述通信信号波形的复制，建立相应的信息的度量，进而建立信息论的第一、第二和第三编码定理，揭示了信息在通信系统中有效和可靠传输的基本规律。但其局限性也在于此，只研究信息信号波形的复制，舍去了信息的内容和信息的信息价值，而信息内容和信息价值是远比通信更复杂的信息活动（如推理、思维和决策）中最重要的因素。在通信以外的许多场合，信息不一定符合概率统计规律。概率熵必须推广到非概率的情形，以便能够有效而统一地描述和度量信息的形式、内容和价值。综合地考虑信息的形式因素（语法信息）、含义因素（语义信息）和效用因素（语用信息），即“全信息”。研究全信息的本质、全信息的度量方法及全信息的运动（变换）规律的理论被称为“全信息理论”。该理论引入主观因素、非形式化的因素和模糊、混沌因素，重视主观与客观相互作用、非形式化和形式化有效结合，强调用新的科学观、新的方法论和新的数学工具研究信息的本质。

关于知识与信息、知识与智能之间的关系问题，他指出“知识是对信息进行加工提炼所获得的抽象化产物，对问题和环境信息进行处理而生成知识，知识被目的激活而生成智能。”那么，如何把信息提炼成为知识呢？由于信息表达的是事物的状态及状态变化的方式，知识表达的是事物运动的状态及状态变化的规律。因此，由信息生成知识的归纳过程本质上就是一个由个别事物运动状态的具体变化“方式”升华为一类事物运动状态的普遍变化“规律”的抽象化过程，即飞跃过程。而知识与智能之间的转化则是通过形成求解问题的策略来实现的，求解问题的策略又是在求解问题的目标的引导下由相关的知识生成（称之为“再生”）的。

钟义信从定性和定量两个方面为研究知识理论建立起必要的基础，进而在此基础上阐明知识理论的核心命题，即“第一，如何通过归纳机制把信息提炼成为“知识”；第二，如何通过目标导引把“知识”激活成为“智能”。这样两个核心命题的初步阐明，在一定程度上揭示了信息、知识、智能三者之间内在的本质联系。

4. 图书情报界的知识论研究

图书情报理论基础中的知识论研究始于20世纪80年代。主要包括知识基础论、知识交流论、知识组织论、知识集合论和知识管理论五个方面。

知识基础论是图书情报知识论研究的基础，它为后来图书情报知识论的研究打下坚实的理论基础，并提供了广阔的思维空间。知识基础论的产生首先来自英国哲学家波普尔（Karl Popper）的“世界3”理论。他认为，世界1是物质世界，世界2是精神世界，世界3是知识世界。知识世界包括理论、问题和论据，它具有真实性（与物质对象一样真实地存在，并作用于物质）、部分自主性（世界3本身可自主产生一个理论）和永恒性（由人心也即世界2创造，并反作用于人心）等特征。世界3理论对图书情报学的研究也产生了重大影响。尤其是他的“客观知识论”的思想为图书馆学、情报学的研究开阔了视野，许多学者甚至把它作为图书馆学、情报学的理论基础。近年来，中国学者对波普尔的世界3理论进行了较深入的

探讨,并结合近年来信息化、网络化、计算机化的实际,对世界3理论进行了修正,用“编码”、“文本”的概念限定世界3的有关表述,以计算机能够做出一些人脑做不出的发现为依据,提出赛伯空间和虚拟现实既不是单纯的世界1,也不是单纯的世界3,它们是一个动态过程的体现,是这两个世界相互作用的体现。对世界3理论的深入研究,对于找出理解信息时代的理论平台,对于建立和完善知识理论体系,具有积极意义。

知识交流论诞生于20世纪80年代,并很快被图书情报学研究工作者普遍接受,成为贯穿这一时期图书情报学研究的主流。它主要研究交流中的知识、交流过程、知识交流与图书馆及图书馆的知识交流机制。

“知识组织”的观念最早由英国著名分类法专家布利斯(H.E.Bliss)于1929年提出,并一直受到图书馆学界的关注。1993年国际性学术期刊《国际分类(International Classification)》改名为《知识组织(Knowledge Organization)》,这更加证明图书馆与知识组织的相关性。关于知识组织的定义,目前尚未统一。布鲁克斯(B.C. Brookes)认为:“知识组织是对文献中所含内容进行分析,找到人们创造与思考的相互影响及联系的节点,像地图一样把它们标示出来(即知识地图),以展示知识的有机结构,为人们直接提供创造时所需要的知识。”而A. Sigel的解释是:“知识组织是将含有知识的集合物加入信息价值的一种跨学科领域的文化活动。”J.D. Anderson则认为:“知识组织是有关文献的描述、内涵、特色、目的及将前述这些活动予以制止,以利于使用者的寻找。知识组织包含了索引、摘要、编目、分类、记录管理、书目及相关文献信息的产生或检索用的书目资料库。”虽然知识有显性知识与隐性知识之分,但在组织时还是以显性的知识为主的。知识整序的方式除了文献的分类、标引、编目、文摘等之外,还需进一步针对更细微的内容部分——知识单元给予揭示和分析。实际上,知识组织比信息组织的内容分析更具深度,且兼顾知识相关性的活动。台湾学者曾对人工智能、认知心理学、语言学、图书馆学、情报学对知识组织的方法进行过比较,认为文献组织有别于知识组织;知识关联性是知识组织的重要特性;知识单元具有分合特性。进而提出今后有关知识组织的研究要结合人工语言与自然语言,以知识单元为基础、重视层面分类理论的应用等建议。

国内学者周文俊的文献交流论、宓浩和黄纯元的知识交流学说、王子舟的知识集合论都是有关研究成果。有的学者认为,知识组织理论可以成为图书馆学的理论基础。在微观上,知识组织理论揭示了图书馆文献组织的实质是知识组织的原理,给图书馆内部活动以知识组织为目标的恰当说明;在宏观上,知识组织理论又对图书馆的产生、组织原理、基本职能、社会定位及对图书馆学的研究范围、基本原理、学科定位等问题,也都能做出合乎实际、合乎逻辑的说明。

知识集合论认为,图书馆的实质就是知识集合,图书馆学的研究对象应转移至知识集合的命题上来。所谓知识集合,即用科学方法把客观知识元素有序地组织起来,形成专门提供知识服务的人工集合,当然,具有知识集合实质的不仅仅是图书馆,还有其他集合现象具有知识集合的实质,如百科全书、知识数据库等。确定知识集合为图书馆学的研究对象,有助于厘清图书馆学的研究范畴,拓宽图书馆学的研究领域,提高图书馆学的学术地位,纯化图书馆学内容体系,并摒除图书馆学中的空疏学风,总而言之,将对图书馆学的发展起到巨大的推动作用。

20世纪90年代以来,随着起源于企业管理的“知识管理”概念的兴起,知识管理理论在图书情报界的研究日益增多,国内有的学者在对相关研究现状进行调研的基础上提出:知识管理理论是图书情报学知识论中的集大成者。由于知识交流、知识组织和知识集合从某种意义上讲均属于知识管理的范畴,因此可以说知识管理理论是图书情报学知识论的核心部分。

2.1.4 本体论

1. 哲学意义上的本体论

“本体论”(Ontology)这个词出现于17世纪,最早见于德意志哲学家郭兰克纽1613年出版的《哲学辞汇》一书。Ontology由“ont”和“ology”组成。“ont”源出希腊文,是“on”(όν)的变式,相当于英文的“Being”,即“存在”;“ology”为“学问”、“学说”,二个词缀何在一起表示“有关存在的学问”。人们一般都把它当做是从柏拉图到黑格尔的西方传统哲学的主干,或“第一哲学”。这意味着它是各个哲学分支的理论基础,是理论中的理论、哲学中的哲学;其他哲学问题都是围绕着建设、运用或怀疑、反对本体论而展开的。

虽然本体论这门学问可以追溯到柏拉图,但是直到18世纪才出现它的定义。第一个为本体论下定义的是德国哲学家沃尔夫(Christian Wolf, 1679—1754),他指出,“本体论,论述各种抽象的、完全普遍的哲学范畴,在这个抽象的形而上学中进一步产生出偶性、实体、因果、现象等范畴。”这个定义表达了本体论作为西方哲学特有的一种哲学形态,其中包含着中国传统哲学中所没有的思想方法。我国《辞海》中关于“本体论”的描述是:“哲学中研究世界的本原或本性的问题的部分。”《中国大百科全书·哲学卷》中指出“本体论在西方哲学史和中国哲学史中分别具有各自的含义,在古希腊罗马哲学中,本体论的研究主要是探究世界的本原或基质,在中国古代哲学中,本体论指关于世界的本源、本体或本根的学说”。对此“本原”的研究即成为本体论的先声,而且逐步逼近于对“Being”(是、存在)的探讨。

在我国,有学者从本体论的定义出发,探讨了本体论所具备的特征,认为本体论具有三个基本的特征:从实质上讲,本体论是与经验世界相分离或先于经验而独立存在的原理系统,这种哲学当然应归入客观唯心主义之列;从方法论上讲,本体论采用的是逻辑的方法,主要是形式逻辑的方法,到了黑格尔发展为辩证逻辑的方法;从形式上讲,本体论是关于“是(存在)”的哲学,“是”是经过哲学家改造以后而成为的一个具有最高、最普遍的逻辑规定性的概念,它包容其余种种作为“所是”的逻辑规定性。因而得以命名,即它是一门关于“是”的学问,其较适当的译名应为“是论”。

2. 本体论在网络信息组织中的应用

本体论是哲学的概念,它是研究存在的本质的哲学问题。本体论在哲学意义上的主要特点在于本体论是关于世界某个方面的一个特定分类体系,这个体系是不依赖于任何特定的语言的。20世纪90年代以来,随着现代科技和信息科学的飞速发展,本体论开始渗透到知识工程、人工智能、信息系统、计算机科学、电子商务、数字图书馆等多个学科领域,并逐渐成为这些研究领域广泛关注的一个重要概念。例如,在知识工程界与信息科学领域,1993年美国斯坦福大学知识系统实验室的Gruber给出了被广泛认可的定义,“Ontology是共享概念模型的明确的形式化规范说明”,“是对概念化的精确描述”。在人工智能界,Neches等人将本体定义为:“构成相关领域词汇的基本术语和关系,以及由这些术语和关系构成的解释这些词汇外延的规则。”

简言之,本体论就是描述概念及概念之间关系的模型或详细说明,通过概念之间的关系来描述概念的语义。一个Ontology往往就是一个正式的词汇表,其核心作用就在于定义某一领域或领域内专业词汇及它们之间的关系。这一系列的基本概念如同一座大厦的基石,为交流各方提供了一个统一的认识。在这一系列概念的支持下,为计算机理解文本含义提供背景知识,使人与计算机之间的交流方便、快捷。

本体论可以分为四种类型:领域、通用、应用和表示。领域本体包含着特定类型领域(如电子、机械、医药、教学)等的相关知识,或者是某个学科、某门课程中的相关知识;通用

本体则覆盖了若干个领域，通常也称为核心本体；应用本体包含特定领域建模所需的全部知识；表示本体不只局限于某个特定的领域，还提供了用于描述事物的实体，如“框架本体”，其中定义了框架、槽的概念。

在实际的应用中，本体论学者、知识管理、人工智能、情报学（图书馆学）甚至任何一个具有大量需要归类和划分信息的部门及领域都可以成为本体论的应用对象。本体论的基本元素是词汇（Term）/概念（Concept），转而构成同质化的类（Class）和子类（Sub-class），然后各个类和概念之间加入了适合的关系（Relation）后，形成了一个简单的本体。概念和类皆用来表达词汇本身，而关系则为词汇提供连接（Mapping），并加入限制条件（Constraint），使之与现实情况相符合。构建本体的简单步骤是：

- （1）列出研究课题所涉及的词条（Terms）；
- （2）按照词条的固有属性和专属特征进行归纳和修改，对词条建立类（Class）及层级化的分类模型（Taxonomy）；
- （3）加入关系（Relation）连接 Terms 和 Taxonomies；
- （4）按照需要，添加实例（Instance）作为概念的具象。

关于本体论在信息组织领域的应用，国内有很多学者分别从不同的角度进行了分析和概括，代表性的观点认为，从信息组织的角度出发，本体及其方法和技术为信息组织，特别是网络信息组织带来了新的变革，这些变革主要表现在直接体现语义的网络信息组织、分布式共享，多维、网状的信息组织方式及对推理的支持上。

首先，由于传统的网络信息组织所表达的语义都是隐含的，不能直接表达为机器（计算机）所理解的形式化语义，而基于本体的网络信息组织不仅方便计算机的“理解和处理”，而且还可以在此基础上提供进一步的智能服务，此外，由于本体在表达信息内容的概念时是在一定的语义环境或限制规则下完成的，因此在表达概念及其含义时更加清晰和准确，在进行信息组织时也更加规范。

其次，在分布式共享方面，对于领域知识的共同理解与描述，并不一定要通过一个集中管理的本体来完成，可以由分散在网络上的多个本体来实现，即本体为实现分布式共享提供了相应的引入机制，这种分布式共享的信息组织方式，不仅可以降低信息组织建立、维护与管理的成本，而且还可以大大促进网络知识的共享与交流。

最后，本体采用了容易为计算机所接收和处理的体现描述逻辑的知识表现和信息组织方式，概念及其之间的关系形成了一个多维的语义网络，这种多维、网状的信息组织方式，不仅有利于网络上各种不同类型、不同结构的信息资源的集合与整合，而且更加有利于它们之间关系的描述和揭示。最后，本体及其所具备的推理能力代表了现代信息组织，特别是网络信息组织的发展趋势，它不仅有利于信息的形式化描述，而且依据本体这种信息组织方法而建立的检索系统，更能满足用户进行语义检索，特别是智能检索。不过，不同的本体有不同的结构。“有些本体是某一狭窄主题领域词汇的分类体系（如分类表或等级体系表）。有些是概念特征集合说明，如元数据方案就是规定使用所用的元素、元素含义及元素属性的类型和属性值的主体。还有一些本体看起来像分类控制词表，就像 WordNet，对词汇进行语义分析，将其分别归入名词、形容词和副词类。这类本体与只给出专业词汇的名词形式（当然还有一些修饰词对名词进行限制）的标题词表和叙词表形成了鲜明对比。”^①。总而言之，本体论为信息组织领域引入了许多新的思想和方法，特别是为网络信息组织带来了新的机遇。

有关本体具体的应用请见本书第3章信息描述语言与编码第3.3节“语义网环境中的形式化描述语言——本体”和第3.4.6节“本体描述语言”。

^① Arlene GTaylor.信息组织. 张素芳, 等译. 北京: 机械工业出版社, 2006: 200.

2.1.5 分形理论

分形理论是当今世界十分风靡和活跃的新理论、新学科。分形的概念是美籍数学家曼德布罗特(B.B.Mandelbort)首先提出的。1967年他在美国权威的《科学》杂志上发表了题为《英国的海岸线有多长?》的著名论文。海岸线作为曲线,其特征是极不规则、极不光滑的,呈现极其蜿蜒复杂的变化。我们不能从形状和结构上区分这部分海岸与那部分海岸有什么本质的不同,这种几乎同样程度的不规则性和复杂性,说明海岸线在形貌上是自相似的,也就是局部形态和整体形态的相似。在没有建筑物或其他东西作为参照物时,在空中拍摄的100公里长的海岸线与放大的10公里长的海岸线的两张照片,看上去会十分相似。事实上,具有自相似性的形态广泛存在于自然界中,如连绵的山川、飘浮的云朵、岩石的断裂口、布朗粒子运动的轨迹、树冠、花菜、大脑皮层等,曼德布罗特把这些部分与整体以某种方式相似的形体称为分形(Fractal)。1975年,他创立了分形几何学(Fractal Geometry)。在此基础上,形成了研究分形性质及其应用的科学,称为分形理论(Fractal Theory)。

自相似原则和迭代生成原则是分形理论的重要原则。它表征分形在通常的几何变换下具有不变性,即标度无关性。由自相似性是从不同尺度的对称出发,也就意味着递归。分形体中的自相似性可以是完全相同,也可以是统计意义上的相似。标准的自相似分形是数学上的抽象,迭代生成无限精细的结构,如科契(Koch)雪花曲线、谢尔宾斯基(Sierpinski)地毯曲线等。这种有规分形只是少数,绝大部分分形是统计意义上的无规分形。

分维,作为分形的定量表征和基本参数,是分形理论的又一重要原则。分维,又称分形维或分数维,通常用分数或带小数点的数表示。长期以来,人们习惯于将点定义为零维、直线为一维、平面为二维、空间为三维,爱因斯坦在相对论中引入时间维,就形成四维时空。对某一问题给予多方面的考虑,可建立高维空间,但都是整数维。在数学上,把欧氏空间的几何对象连续地拉伸、压缩、扭曲,维数也不变,这就是拓扑维数。然而,这种传统的维数观受到了挑战。曼德布罗特曾描述过一个绳球的维数:从很远的距离观察这个绳球,可看做一点(零维);从较近的距离观察,它充满了一个球形空间(三维);再近一些,就看到了绳子(一维);再向微观深入,绳子又变成了三维的柱,三维的柱又可分解成一维的纤维。

显然,并没有绳球从三维对象变成一维对象的确切界限。数学家豪斯道夫(Hausdorff)于1919年提出了连续空间的概念,也就是说空间维数是可以连续变化的,它可以是整数也可以是分数,称为豪斯道夫维数。曼德布罗特也把分形定义为豪斯道夫维数大于或等于拓扑维数的集合。英国的海岸线为什么测不准?因为欧氏一维测度与海岸线的维数不一致。根据曼德布罗特的计算,英国海岸线的维数为1.26。有了分维,海岸线的长度就确定了。

分形理论既是非线性科学的前沿和重要分支,又是一门新兴的横断学科。作为一种方法论和认识论,其启示是多方面的:一是分形整体与局部形态的相似,启发人们通过认识部分来认识整体,从有限中认识无限;二是分形揭示了介于整体与部分、有序与无序、复杂与简单之间的新形态、新秩序;三是分形从一特定层面揭示了世界普遍联系和统一的图景。显然,分形理论对信息组织亦有很大的启迪作用。

2.2 信息组织的方法基础

2.2.1 语言学

1. 语言学原理概述

语言学(Linguistic)是指研究语言现象及其本质、特点、结构、功能、起源和发展规律

的科学。一般包括句法学、构词学、音系学、语法学、语用学、语义学等。语言又可被简易地划分为自然语言与人工语言，自然语言是人们日常生活中彼此交流的主要工具，也是人类表达思考方式的工具。人工语言又称人造语言、国际辅助语言等，是为了特定的需要、目的、用途而人为创造的语言。不论是自然语言还是人工语言，都是表达信息、组织信息的最基本的工具。因此，语言学的知识是信息组织的重要基础知识之一。

2. 语言学原理在信息组织中的应用（检索语言等）

一般情况下，信息组织必须借助于某种检索语言或信息检索规则，检索语言是指人们根据信息检索的需要，依据一定的规则对自然语言进行事先规范和控制而编制的一种语言，是信息的描述、存储、分类和信息组织的工具。从“规范”和“控制”的角度看，它属于一种人工语言。

1) 检索语言的类型

人们通常把一部分分类法或一部主题词表称为一种检索语言，目前全世界的检索语言至少有几千种，对检索语言类型的划分方法也有许多种。最常见的是按结构原理来划分，检索语言可分为分类语言、主题语言和代码语言三大类型。

(1) 分类语言，习惯上称分类法，用分类号来表示概念及其在系统中的位置，将各种概念按其所属的学科性质进行分类和排列，是一种按照学科范畴划分而构成的语言体系。分类语言成为反映科学知识分类体系的逻辑系统。世界上著名的分类法有《杜威十进制分类法》、《国际十进制分类法》、《美国国会图书馆分类法》、《国际专利分类法》、《中国图书馆分类法》等，这些分类法在文献信息的加工、组织和检索中发挥着不可或缺的作用，尤其在传统图书馆环境中对于文献信息资料的排架整理功不可没。

分类语言包括等级体系式分类语言、分面组配式分类语言和等级—组配式分类语言。体系分类法是一种直接体现知识分类的等级制概念标识系统，主要特点是按学科、专业集中信息，并从知识分类角度揭示各类信息在内容上的区别和联系，提供一种检索的途径。体系分类法的主要优点是它分类结构直观，易于掌握使用，具有较好的族性检索功能，最大缺点是不能适应于按专指概念进行检索，无法根据现代科学的发展及时自动生成新类，难以与科学的发展保持同步。组配分类法是在体系分类法基础上发展起来的，它克服了体系分类法列举式类目不能容纳主题概念发展的局限性，对科学发展的适应性强，但它不如体系分类法直观，标记复杂，不易掌握。等级—组配式分类法是上述两种分类法的融合，是一种在详尽类表的基础上，广泛采用各种组配方式的分类法。它兼有两种分类法的优点，但标记复杂，类目之间的组配往往需要多种辅助符号。

(2) 主题语言是用词语来表达各种概念，将各种概念按字顺排列的检索语言。主要包括标题词语言（标题法）、单元词语言（单元词法）、叙词语言（叙词法）和关键词语言（关键词法）。亦可统称为主题法系统。

标题词语言是一种先组式的、按标题字顺排列的检索语言。由于先组式语言的局限性，标题词语言逐渐被后组式的主题词语言取代。

单元词语言是一种后组式的检索语言。单元词语言曾经流行一时，后因表达主题时语义不定性而被叙词语言所取代。

叙词语言综合了分类语言、标题词语言、单元词语言的优点，发展迅速。叙词语言以概念组配为基础，更能很好地适应计算机检索系统。

关键词语言是直接以文献信息中能够表达信息主题概念的语词作为标识的一种信息描述语言，是非规范化的主题语言。

(3) 代码语言是一般只就事物的某一方面特征，用某种代码系统来加以标引和排列事物

概念,从而提供检索的信息描述。这种检索语言比较适用于某一专业的检索,如化合物的分子式索引系统、档案号等。

2) 语言学原理在检索语言中的作用

现代语言学经历了历史比较语言学、结构主义语言学和转换生成语言学三个阶段,已发展成一个庞大的学科体系,包括普通语言学、计算语言学、数理语言学、应用语言学等,这些学科分支和信息组织中的检索语言有着或多或少的联系。

(1) 普通语言学的应用。

检索语言的发展和应用与普通语言学有着密切的关系。从20世纪中叶开始,西方学者开始把语言学的一些原理运用到检索语言中。被称为信息检索“第二次革命”的陶布单元词语言,就根据美国描写语言学的“分布理论”制定了单元词的“同现关系”。赛格尔将“信息检索语言”、“文献工作语言”这些概念代之以“信息语言”(Information Language),也是希望对其加深语言学的阐释。而斯巴克·琼斯和马丁·凯的《语言学与情报学》(Linguistics and Information Science)及哈钦斯的《标引和分类语言——结构和功能的语言学研究》(Languages of Indexing and Classification—a Linguistic Study of Structures and Functions)则是我们现今能读到的关于情报检索语言最全面的语言学研究著作。我国学者对语言学在检索语言研究中的应用也已有了基本认识。张琪玉的《情报检索语言语法体系初探》就是我国第一篇用语言学研究体系来探讨检索语言体系结构的论文^①。

(2) 计算语言学的应用。

计算语言学指的是这样一门学科,它通过建立形式化的数学模型,来分析、处理自然语言,并在计算机上用程序来实现分析和处理的过程,从而达到以机器来模拟人的部分乃至全部语言能力的目的。计算语言学应用于语音合成、语音识别、信息检索、信息抽取、人机接口、机器翻译等领域。

计算语言学在信息组织中的应用,如运用词频统计分析方法进行标引和检索词汇的选择。在语言学中运用词频统计分析是定量研究的传统内容。这些统计分析对检索语言的词汇选择有重要意义。例如,在编表时,先收集原始文献中的全部术语,统计其词频,研究其分布特征,排除没有检索意义的高频词和低频词,最后确定适当频率的词编入词表。此外,还可对文献中词汇频率进行计算,以此进行抽词标引或自动分类。例如,逆文献加权标引就通过统计某词在文献中出现的次数(词出现频率)和包含某词的文献数(词文献频率)来选择标引词,即选择那些词出现频率较高、词文献频率较低的词,并根据词出现频率设计权值^②。

2.2.2 逻辑学

1. 逻辑学原理概述

逻辑学是一门具有悠久历史、关于科学认识和思维方法的科学,包括形式逻辑、数理逻辑、辩证逻辑等一系列逻辑分支。其中,形式逻辑研究思维的逻辑形式和基本规律,是逻辑科学体系中最基础的知识。思维主要包括形象思维和抽象思维两种:感觉、知觉和表象是形象思维形式,属于感性认识的阶段;概念、判断和推理是抽象思维形式,属于理性认识的阶段。信息组织是一种智力活动,离不开人的逻辑思维,也就是说,进行信息组织加工必然用到形式逻辑的一些方法,信息组织的行为只有符合逻辑思维规律,才能保证信息组织的序化质量。因此,信息组织遵循科学的思维方法,是在各种概念的基础上进行的。实践表明,信

^① 戴维民. 信息组织. 北京: 高等教育出版社, 2004:18~19.

^② 戴维民. 信息组织. 北京: 高等教育出版社, 2004:19~20.

息组织者的逻辑思维能力越强，其信息组织工作也越好。事实上，信息组织是对各个信息对象，经过从事物到概念再到语言的层层递进的分析之后，运用科学思维使之序化的一个过程。

2. 逻辑学原理在信息组织中的应用

1) 概念的内涵和外延

按照形式逻辑，一个概念一般都包括内涵和外延两个方面。内涵是指所有组成该概念的事物的本质属性，外延是指具有这种本质属性的所有事物的总和。例如，“图书分类法”这一概念，它的内涵是指一切用于图书分类的分类法的总称，它的外延则是所有符合这一含义的分类法，如“中国图书馆分类法”、“杜威十进制分类法”、“美国国会图书馆分类法”等。概念的内涵和外延之间是此消彼长的关系，内涵扩大，则外延缩小；内涵缩小，则外延扩大。如图书馆、公共图书馆、县级公共图书馆，这三个概念的内涵是不断扩大的，但其外延是不断缩小的，如图 2.1 所示。

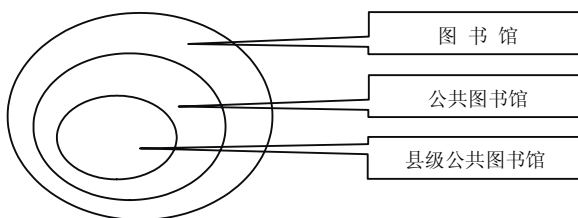


图 2.1 概念的内涵和外延

在信息组织中，利用概念的内涵和外延对各种信息进行标引和检索，依据概念的限制和概括及概念之间的关系，通过概念的划分、概念的综合开展信息组织工作。概念的划分、分析和综合是信息组织中普遍使用的方法，对概念的划分可以建立等级性的概念体系，对概念的综合，则可以将复合主题概念分解成若干个子概念，并通过对子概念的组配表达复合主题概念，进行概念逻辑运算。布尔逻辑示意图如图 2.2 所示。

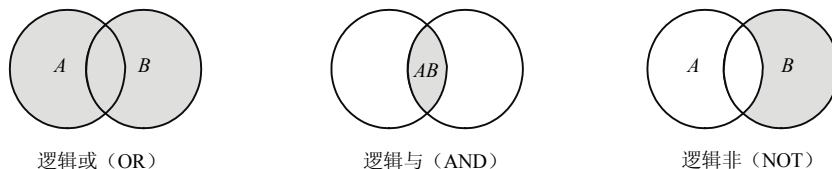


图 2.2 布尔逻辑示意图

2) 概念之间的关系

概念之间按照是否存在共有的外延，概念关系可以分为相容关系和不相容关系两种，如图 2.3 所示。

(1) 相容关系是指至少有一部分外延相同的概念之间的关系，包括同一关系、包含关系、交叉关系等。

① 同一关系，是指具有相同外延的概念之间的关系。如自行车和脚踏车、西红柿和番茄等。同一关系的概念的外延完全相同，可以在标引和检索时相互代替，其逻辑关系可以表示为 $A=B$ 。

② 包含关系，又称属种关系，指一个概念在另一个概念的外延之中，并且是另一个概念的组成部分。如经济和农业经济、图书馆和高校图书馆。在包含关系概念中，包含的概念称为上位概念或属概念，被包含的概念称为下位概念或种概念。包含关系的特点是上位概念

的属性一般存在于下位概念之中,因此二者在外延和内涵上存在密切关系,其逻辑关系可以表示为 $A \supset B$ 。与包含关系类似的还有整部关系和方面关系:整部关系是指整体和部分的关系,如人和手;方面关系为整体概念与方面概念之间的关系,如人和体重。二者都不属于包含关系,其概念属性之间的联系不如包含关系密切。

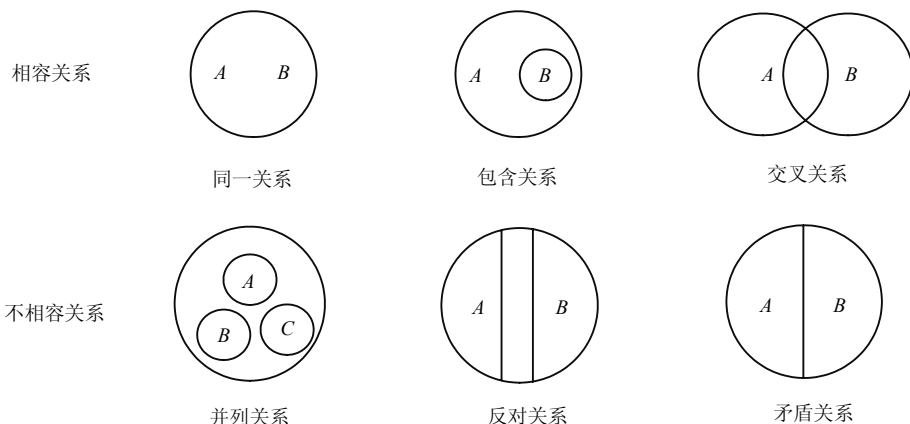


图 2.3 概念之间的关系类型

③ 交叉关系,指部分外延重合的概念之间的关系。如妇女和医生、中国人和工程师。交叉关系只是部分外延重合,其重合部分构成一个新的概念,它具有二者的属性,如女医生、中国工程师。其逻辑关系可以表示为 $A \cap B \neq \emptyset$ 。

(2) 不相容关系是指不存在共有外延的概念之间的关系,可以分为不同论域之间的不相容关系和同一论域的不相容关系。前者对于信息组织而言无实际意义,信息组织关注的是同一论域的不相容关系,可以分为并列关系、反对关系和矛盾关系。

① 并列关系,指一个上位概念下几个不存在共有外延的并列下位概念之间的关系。所有并列概念之和等于上位概念,并列概念之间是相互排斥的。其逻辑关系可以表示为 $A \cap B \cap C \dots = \emptyset$, 且 $A \cup B \cup C \dots = I$, I 为其并列概念的上位概念。如“亚洲国家”可以分出“中国”、“韩国”、“日本”等。

② 反对关系,指外延之和小于上位概念外延的两个不相容概念之间的关系,而二者不具有共有的外延。其逻辑关系可以表示为 $A \cap B = \emptyset$, 且 $A \cup B \subset I$, I 为其二者的上位概念。如“公共图书馆”和“高校图书馆”,二者之和小于“图书馆”,但它们又不存在共有的外延。

③ 矛盾关系,指外延之和等于上位概念外延的两个不相容概念之间的关系,其特点是二者之和等于上位概念,而二者又不存在共有的外延。其逻辑关系可以表示为 $A \cap B = \emptyset$, 且 $A \cup B = I$, I 为二者的上位概念。如“核国家”和“非核国家”,二者之和等于上位概念“国家”的全部外延,二者本身又不存在共有的外延。

3) 概念的划分

概念的划分是指以对象一定属性为标准,将一个属概念的外延分成若干种概念以明确其外延的逻辑方法。一类事物,有的由一个或几个事物组成,而有的则由许多甚至是无法计数的事物组成,不可能也没必要把它们一一列举出来。例如,我们不可能列举宇宙中的每颗星星,但可以用星星自身是否发光作为标准,可以把星星分为恒星和行星两大类。这里应引起注意的是划分与分解的区别:划分是把一个属概念分为若干个种概念,分解是把一个具体的事物肢解成许多构成部分。

划分由母项、子项和划分根据三个部分构成。划分的母项就是被划分的概念,也是属概

念；划分的子项就是划分概念，也是种概念；划分根据就是划分时采用的某一标准。因此，划分就是把属概念分为若干个种概念的逻辑方法。例如，依照人的性别，把人分为男人和女人，那么人就是划分的母项，男人和女人就是划分的子项，性别就是划分根据。

（1）划分的方法。

根据划分的层次可以分为一次划分（一层）和连续划分（多层）。此外，还有一种特殊的划分，叫二分法。

一次划分，就是只对母项进行一次划分，后面不再划分。例如，把人划分为男人和女人，以下不再进行划分。

连续划分，就是经过一次划分后，再把子项当成母项继续进行划分，直到满足需求为止。例如，把人分为男人和女人后，又把男人分为50岁以上的男人和50岁以下的男人等。

二分法是根据一个概念中的矛盾关系进行划分，把属概念划分成两个互相矛盾的种概念。二分法可以是一次划分，也可以是连续划分。

（2）划分的规则。

① 划分必须是相应相称的。

就是说，划分的子项和划分的母项应该是对应的，各子项外延之和必须和母项的外延一致，不能多出子项，也不能少了子项。违反这条规则所犯的逻辑错误，或是“划分不全”，或是“多出子项”。例如，对于群众的思想状况，应当划分为先进、中间、落后三种，多了没有必要，少了不符合实际。有时，当划分暂时不需要或还无法列举所有的子项时，可以列出主要项目，其余的用“其他”来概括。

② 划分出的子项必须是互相排斥的。

就是说，划分出来的子项必须是不相容的并列关系，不能是包含关系或交叉关系，否则会犯“子项相容”的逻辑错误，不能起到划分的作用。例如，如果把入划分成男人、女人、中国人和外国人，男人和女人这两个种概念与中国人和外国人这两个种概念就是交叉的，而不是互相排斥的，这样的划分就达不到划分的目的。

③ 每次划分必须按同一标准进行。

对同一概念的划分，可以根据不同的需要，选择事物的不同属性作为划分根据，进行多种不同的划分。连续划分时，每次划分的标准也可以不同。但是同一次划分，只能选择同一个划分标准，违反这条规则会犯“标准不一”的逻辑错误。例如，我们可以按照性别把人划分为男人和女人，可以按照国籍把人划分为中国人和外国人，但是不可以同时按照性别和国籍两个标准，把人划分成男人、女人、中国人和外国人，这和前面的标准“划分出的子项必须互相排斥”中所举的例子是一样的道理。

划分的三条原则，既有区别，又密切相关。一个划分，如果违反了一条规则，往往也违反了其他两条规则。当然，除了掌握划分方法和划分规则，对划分对象具体知识的了解也是必需的。一个对化学知识一点不懂的人，即使掌握了划分方法和规则，也不能对化学这个概念做出正确的划分。

显然，按形式逻辑将学科划分成线形的上下位类或同位类关系，在管理、研究上有其方便之处，在许多情况下是合理和必要的，但也应清醒看到，在一定的条件下，其局限性也十分明显：因为它有可能把本来相互有密切联系的类型隔裂开来了。尤其当人们进入复杂的研究领域时就更是如此。因此，在某些情况下，一定不能拘泥于形式逻辑的分类，不能一味地“非此即彼”，而应该根据客观的情况，要“亦此亦彼”，即要尽可能反映事物的本来面目。当已有的概念、类目名称不能反映实际情况时，应大胆地根据实际情况概括出新概念和类名；当根据研究工作的需要，以问题或主题为中心、为导向时，就需要打破学科界限，利用辩证

思维,构建更接近客观世界的学科体系。这里的关键是一定要注意形式逻辑和辩证逻辑的适用范围和一定的研究目的,既不能不顾条件盲目地使用,又不能将两者对立起来。

近年来,由于网络资源的发展,出现了要根据不同的目的和需求构建新的学科体系,需要引进新的思维方式,将形式逻辑与辩证逻辑有机结合起来的问题。目前多数分类体系是线形性的、层层隶属的展示方式,但有的则是非线形的树状方式,有的则是将线形与非线形有机结合在一起的方式。目前文献分类已出现将学科分类与主题结合的分类体系——《中国分类主题词表》,该表是在《中国图书分类法》和《汉语主题词表》的基础上编制的、两者兼容的一体化信息检索语言,是一部大型综合性的信息文献标引工具。目前的网络综合性分类搜索引擎多数采用以主题对象为中心的组织方式,即按事物对象和学科分别设置类目,可以集中一个主题对象的有关信息资源,必要时以事物为中心建立两者之间的联系。这种不同于传统分类法的组织方式是按照主题为中心展开的体系,在类目体系纵向展开上,使用了多维划分和多元展开的形式。对事物对象的揭示形式直观、直接性好,比较符合普通用户按对象和问题检索的习惯,目前已经成为一种在网络上占主导地位的实用组织模式^①。这种以主题为中心的分类结构可以引起思维方式上的思考。在面对学科分类一般问题时,形式逻辑的思维方式和各项规则无疑还是有效的。但面对复杂的问题和实践中出现的新问题,当应用形式逻辑的方法不能解决时,可以大胆地采用辩证思维方法(辩证逻辑),突破传统的规则,只要分类能够合理地解释问题和解决问题即可。

2.2.3 知识分类学

1. 哲学家的知识分类

科学分类体系提出对科学整体关系的了解,涉及对世界的看法,是一个哲学问题,古今中外许多哲学家都对知识进行分类。亚里士多德、培根、以狄德罗为首的法国百科全书派、牛顿与林耐学派、圣西门和黑格尔都曾经投身于这项工作。19世纪50~70年代,马克思曾预言今后自然科学与社会科学将逐步融合成一门科学,并认为数学化程度是衡量一学科成熟与否的标志。恩格斯在《自然辩证法》一书中,将科学分类问题与物质运动形式联系起来考察,批判了以往科学分类中的机械唯物主义与唯心主义观点,确立了辩证唯物主义的科学分类原则与发展原则。20世纪中叶,毛泽东根据科学对象所具有的特殊的矛盾性,将所有知识区分为自然科学、社会科学两大类,哲学则是两大类知识的概括与总结。20世纪60年代以来,科学发展突飞猛进,人们发现了许多新的物质形态和运动形式,一方面,现代科学高度分化,产生出一系列新的学科;另一方面又高度综合,各门学科相互渗透,不断产生出新的边缘学科、横断学科(研究对象不是客观世界的某种物质结构及其运动形式的某个共同方面,其概念和方法在各门科学中都具有普遍的适用性和方法论意义)和综合性学科。同时,科学的数学化、致用化等趋势为科学分类提出了新的课题。一些哲学家、科学家也提出过一些新见解。

1) 亚里士多德的知识分类

亚里士多德(公元前384—公元前322)是古代希腊哲学家。他从人类活动出发,根据当时社会上业已存在的某种知识分工,将人类的知识分为三大类:第一类为纯属认识活动的学问,称为理论哲学,包括逻辑学、物理学、数学、形而上学等;第二类为研究人的行为的哲学,称为实践哲学,包括伦理学、经济学、政治学等;第三类为关于创作、艺术的学问,

^① 马张华. 分类搜索引擎对分类法发展的贡献及相关问题讨论. <http://www.cnindex.fudan.edu.cn/zgsy/2005n3/mazhanghua.htm>.

称为创造哲学，包括诗歌、艺术等。

2) 培根的知识分类

培根（1561—1620）是英国有名的哲学家，是名副其实的近代科学思想的先驱，他首先对近代早期发展的自然科学做了系统的描述。按照培根的观点，人的学术起源于理解力的三种官能——记忆（Memory）、想象（Imagination）和理性（Reason）。他以此为基础开始了他对知识的分析和分类。记忆对应历史，而历史包括自然史和文明史，二者之下进而各有细分。想象对应诗歌，诗歌分为叙事诗、戏剧诗和寓言诗。理性对应哲学或科学，哲学划分为三种：人类哲学、自然哲学和神学。培根的分类大纲如下。

历史（History）

自然史（Natural History）

文明史（Civil History）

诗歌（Poetry）

叙事诗（Narrative）

戏剧诗（Dramatic）

寓言诗（Paraboliical）

哲学（Philosophy）

人类哲学（Human Philosophy）

文化哲学（Civil Philosophy）

神学（Theology）

培根的分类没有在现象的世界和非实在的形而上学思维的产物之间划出明确的区分，而且学科用语中有中世纪神学的残迹和经院哲学的弊病，因而从近代科学的立场来看是有缺陷的。但是，培根指出：“知识的划分不像以一个角度相交的几条线，而更像在一个树干上交叉的树枝。”这个观念明显区别于以神学为首位的知识体系，比较客观地解释了人类知识现象，在当时具有积极的作用。1870年，英国哲学家和教育家哈里斯首次提出倒转培根知识体系来编制图书分类法，即“倒转培根分类法”，倒转后的分类变成哲学、诗歌、历史。1876年，杜威在哈里斯分类法的基础上编制了杜威十进制分类法。

3) 恩格斯的知识分类

恩格斯（1820—1895）是马克思主义的创始人之一。19世纪70年代，恩格斯在《自然辩证法》一书中，批判了以往科学分类中的机械唯物主义与唯心主义观点，确立了辩证唯物主义的科学分类原则。恩格斯认为，各种不同的物质运动形式是科学研究的主要对象，每门科学所研究的或者是个别的运动形式，或者是相互关联、相互转化的运动形式。因此，科学的划分必须依据运动形式的划分。各门科学之间的相互联系也取决于各种不同的物质运动形式的关系，各门科学在历史上的发展顺序是物质运动形式演化的反映。恩格斯依据当时的科学材料，把多种多样的物质运动形式概括为机械的、物理的、化学的、生物的和社会的五种基本运动形式。他第一次提出了科学分类的两条原则，即客观原则和发展原则，创立了按物质运动形式进行科学分类的正确理论。根据这些原则，恩格斯把科学分为：数学、力学、天文学、地质学、物理学、化学、生物学、社会科学和关于人类思维规律的科学。恩格斯还特别注意研究与各种运动形式之间的转化相联系的各邻近学科之间的转化问题，并预言在这些被忽略的边缘领域有望取得最大的成果，而生长出新的学科。

4) 毛泽东的知识分类

毛泽东在《整顿党的作风》一文中说道：“什么是知识？自从有阶级的社会存在以来，世界上的知识只有两门，一门叫做生产斗争知识，一门叫做阶级斗争知识。自然科学、社会

科学就是这两门知识的结晶,哲学则是自然知识和社会知识的概括和总结。”以《中国图书馆分类法》为代表的我国新型综合型分类法等采用了毛泽东关于知识的概括和分类,建立了五大部类的分类体系^①。

5) 钱学森的知识分类

在20世纪80~90年代,钱学森先生根据研究问题或看问题的角度不同(不是各学科研究对象之不同),将所有知识分成十一大科学技术部门:自然科学、社会科学、数学科学、系统科学、思维科学、人体科学、军事科学、行为科学、地理科学、建筑科学及文艺理论。这十一大类为第一层次,第二层次是“桥梁”,即自然辩证法、历史唯物主义、数学哲学、系统论、认识论、人天观、美学、军事哲学、社会论、地理哲学,建筑科学的建筑哲学。这个“桥梁”分别概括了十一大科学技术部门中带有普遍性、原则性、规律性的东西,即各门科学技术的哲学。第三个层次是马克思主义哲学,其核心是辩证唯物主义。从每一个科学技术部门是直接还是比较间接改造客观世界上划分,所有学科又可以分为基础科学、技术科学和工程技术三个层次(文艺理论目前看来只有一个基础理论层次)。按照钱学森的观点,在现代科学技术十一大部门之外,尚有未形成科学体系的实践经验的知识库,以及广泛的、大量成文或不成文的实际感受,如局部的经验、专家的判断、行家的手艺、文艺人的艺术、中医医药学等,也都是人类对世界认识的珍宝,它与科学技术体系和马克思主义哲学密切相关,不可忽视,亦应逐步纳入体系^②。

钱学森先生的科学分类思想和方法引起了学界兴趣和深入探讨。2008年5月27日~29日,香山科学会议在北京香山饭店召开了主题为“现代科学技术体系总体框架探索”的第324次学术讨论会,来自全国相关单位系统科学、地理科学、思维科学、军事科学、建筑科学、自然科学和社会科学等领域的40多位专家学者应邀参加了会议。与会专家在探讨钱学森现代科学技术体系思想的产生、形成、发展及其科学意义的基础上,重点对系统科学、地理科学和思维科学等领域的研究进展进行了交流,并就存在的主要问题和进一步推进我国现代科学技术体系建设的总体思路、方法和运行管理机制等方面提出了建议^③。显然,尽管有许多问题有待深入讨论,但钱学森先生力图继承和发扬中国古代哲学的精华——整体观,汲取现代科学技术特别是系统科学最新成果,构建新的学科体系总体框架的努力是非常有价值的。

2. 学科的知识分类

学科分类一般根据学科发展的现状,依据各学科研究的方法和对象之间的关系建立。不少国家根据学科的现状建立了自己的学科分类体系。

《联合国教科文组织分类法》是联合国教科文组织建立的分类法,按大学学科分类,其学科体系示意图如图2.4所示。

我国1992年编制的国家标准《学科分类与代码》(GB/T 13745—1992),根据科学性、实用性、简明性、兼容性、扩延性、唯一性的原则,建立了5个门类,门类下分有58个一级学科、573个二级学科、近6000个三级学科,学科体系示意图如图2.5所示。

该标准于2009年进行了修订,同年11月1日正式实施,并更名为《中华人民共和国学科分类与代码国家标准》(GB/T 13745-2009)。学科分类体系的目的是直接为科技政策和科技发展规划以及科研项目、科研成果统计和管理服务的,因此主要收录已经形成的学科,而对于成熟度不够、或者尚在酝酿发展的学科雏形则暂不收录。

① 毛泽东. 整顿党的作风,《毛泽东选集 第三卷》. 北京:人民出版社,2003.

② 逸馥. 钱学森的科学思想与科学成就. 科学时报,2006-12-11(-A2)。

③ 现代科学技术体系总体框架探索——香山科学会议第324次学术讨论会综述. <http://www.xssc.ac.cn/Web/ListConfs/ConfBrief.asp?mo=1194/2008-08-10>.

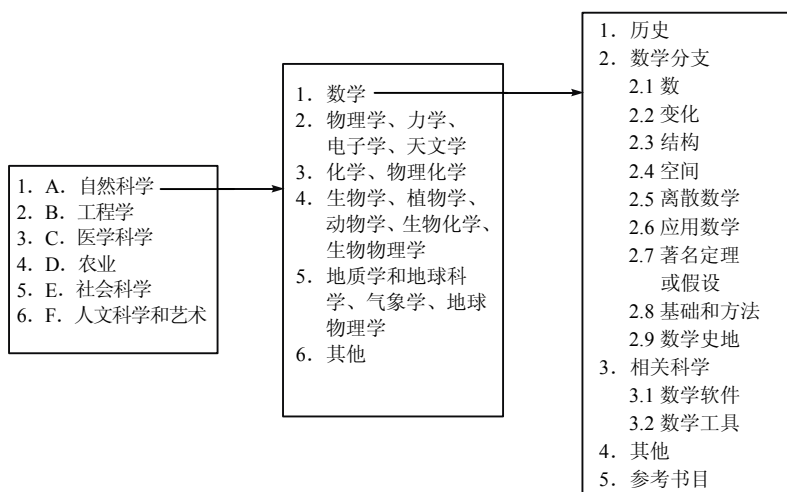


图 2.4 《联合国教科文组织分类法》学科体系示意图

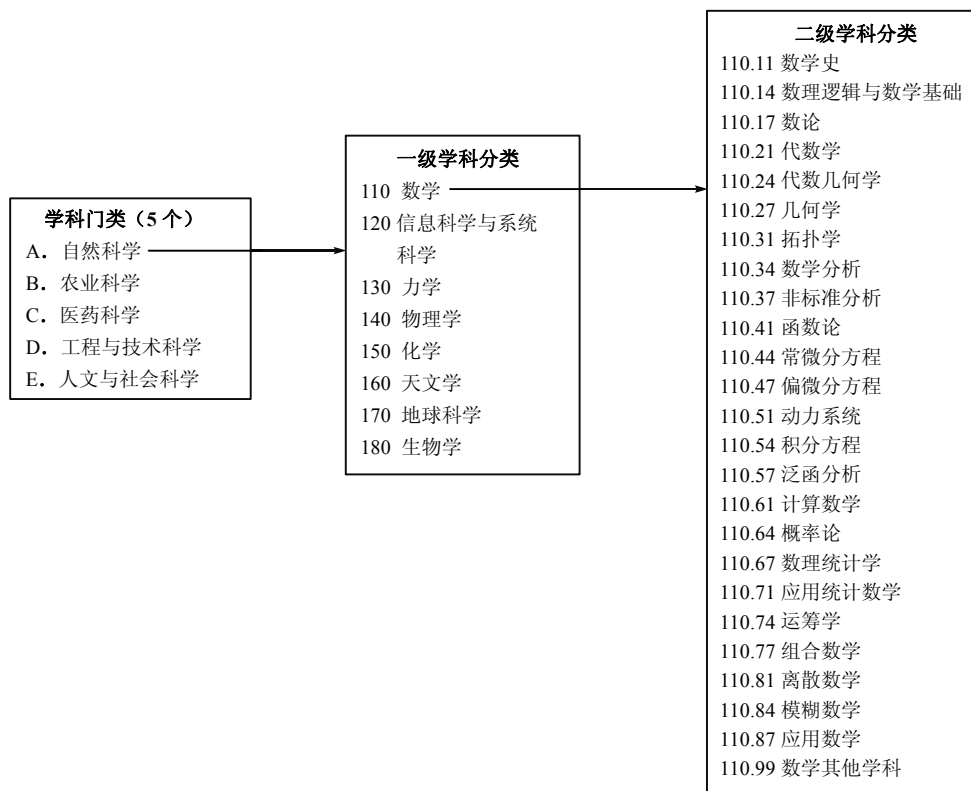


图 2.5 我国《学科分类与代码》学科体系示意图

我国教育部的《普通高等学校本科专业目录(1998 年颁布)》分设哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、管理学 11 个学科门类(无军事学)。下设二级类 71 个, 专业 249 种, 学科体系示意图如图 2.6 所示。

专业目录规定了专业划分标准、专业名称及门类归属, 是设置和调整专业、实施人才培

养、安排招生、授予学位、指导就业、进行教育统计和人才需求预测等工作的重要依据。目录分为学科门类、专业类和专业三级。该目录于2012年修订，同年9月14日发布，2013年正式实行。按照教育部规定，目录十年修订一次；基本专业五年调整一次，特设专业每年动态调整。

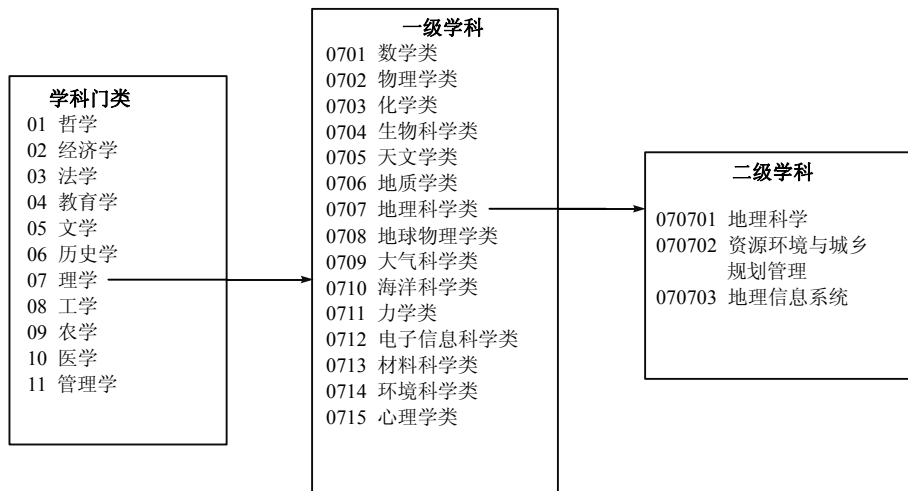


图 2.6 《普通高等学校本科专业目录（1998 年颁布）》学科体系示意图

3. 检索语言下的知识分类

对于检索语言的含义、类型划分、作用等在前文已介绍过，本小节主要介绍《中国图书馆分类法》和《杜威十进制分类法》这两种检索语言的体系结构。

1) 中国图书馆分类法

我国的《中国图书馆分类法》（以下简称《中图法》）是我国图书分类法的基础，《中图法》把一切知识门类按“五分法”分为马列、毛泽东思想、邓小平理论；哲学；社会科学；自然科学；综合性图书等五大部类，在此基础上建成由 22 个大类组成的体系系列。体系示意图如图 2.7 所示。

图书分类法是以科学分类为基础，结合图书资料的内容和特点，对图书进行系统分类的图书分类方法。《中国图书馆分类法》历经多次修订，于 2010 年 9 月出版第五版。

2) 杜威十进制分类法

《杜威十进制分类法》是由美国图书馆专家麦尔威·杜威编制的，对世界图书馆分类学有相当大的影响，已翻译成西班牙文、中文、法文、挪威文、土耳其文、日文、僧伽罗文、葡萄牙文、泰文等出版，并被许多英语国家的大多数图书馆及使用其他相应译文的国家的部分图书馆采用。在美国，几乎所有公共图书馆和学校图书馆都采用这种分类法。

杜威十进制图书分类法于 1876 年首次发表，历经很多次的大改版后，内容已有相当程度的修改与扩充。该分类法以三位数字代表分类码，共可分为 10 个大分类、100 个中分类及 1000 个小分类。除了三位数分类外，一般还会有两位数字的附加码，以代表不同的地区、时间、材料或其他特性的论述，分类码与附加码之间则以小数点“.” 隔开。例如，330 表示经济学，+.9 表示地区，+.04 表示欧洲，即 330.94 表示欧洲经济学。其体系示意图如图 2.8 所示。

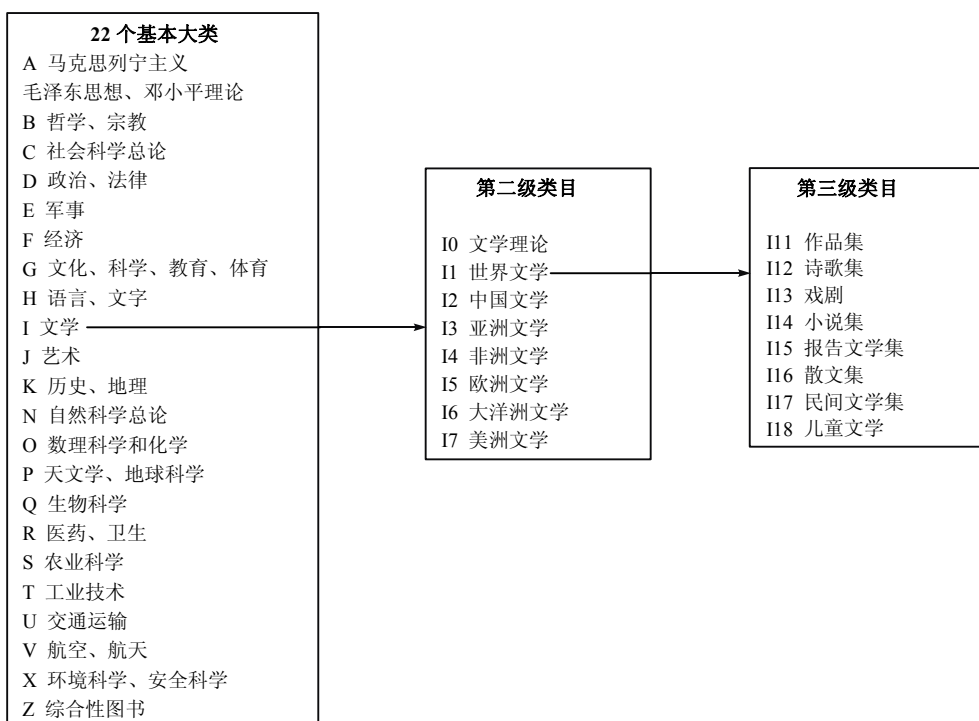


图 2.7 《中国图书馆分类法》体系示意图

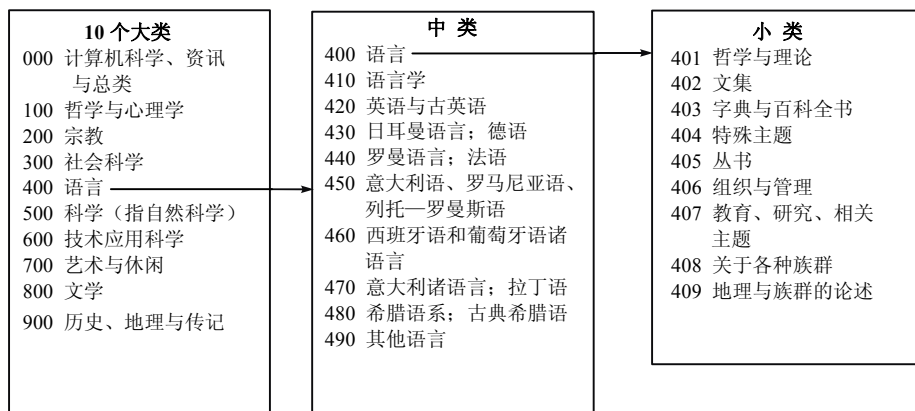


图 2.8 《杜威十进制分类法》体系示意图

4. 现存国内各知识分类的体系比较

目前中国具有实用性、可操作性、最具影响的学科体系应是国家技术监督局发布的《中华人民共和国学科分类与代码国家标准》（GB/T 13745-2009，简称《学科分类与代码》），因为它是“国家标准”，所以对各种有不同需求和目的的学科体系、目录都有指导与参考作用。例如，用于学位教育的教育部和国务院学位委员会颁布《授予博士、硕士学位和培养研究生的学科、专业目录》，以及用于社科研究项目申报的国家社会科学基金委的学科分类表，即国家社会科学基金项目申报数据代码表和用于图书资料分类的《中国图书馆图书分类法》都是以

该标准为基础, 结合各自的特点制定的。《学科分类与代码》的分类对象是学科, 其主要需求和目的是为了便于国家对科研进行宏观管理和统计。为节省篇幅, 下面对几个分类表中有关人文社会科学的类目进行比较。《学科分类与代码》中人文社会科学门类下共有 19 个一级学科 (630-管理学被纳入工程与技术科学门类, 如将管理学纳入人文社会科学, 则一级学科为 20 个), 242 个二级学科。《学科分类与代码》与其他三个文科分类体系对照表如表 2.1 所示。

表 2.1 国家标准学科分类与其他三个文科分类体系对照表

《学科分类与代码》	基金委学科分类	研究生学位学科分类	中国图书馆图书分类法
710-马克思主义	1 马列·科社	8 马克思主义理论	A 马克思主义、列宁主义、毛泽东思想、邓小平理论
720-哲学	3 哲学	1 哲学	B 哲学
730-宗教学	16 宗教学		
740-语言学	19 语言学		H 语言、文字
750-文学	17 中国文学 18 外国文学	12 中国语言文学 13 外国语言文学	I 文学
760-艺术学	艺术学 (暂未列)	15 艺术学	J 艺术
770-历史学	13 中国历史 14 世界历史	16 历史学	K 历史、地理
780-考古学	15 考古学		
790-经济学	4 理论经济 5 应用经济	24 理论经济 3 应用经济	F 经济
810-政治学	7 政治学	5 政治学	D 政治、法律
820-法学	8 法学	4 法学	
830-军事学	军事学 (暂未列)	17 军事学	E 军事
840-社会学	9 社会学	6 社会学	
850-民族学	11 民族问题研究	7 民族学	
860-新闻学与传播学	20 新闻学	14 新闻传播学	在文化、科学、教育、体育类下
870-图书馆、情报与文献学	21 图书馆、情报与文献学	18 图书馆、情报与档案管理 (在管理学下)	在文化、科学、教育、体育类下
880-教育学	教育学 (暂未列)	9 教育学	G 文化、科学、教育、体育
890-体育科学	22 体育学	11 体育学	
910-统计学	6 统计学		
630-管理学	部分入政治类等	17 管理科学与工程	
	2 党史·党建	10 心理学	C 社会科学总论
	10 人口学		X 环境科学、劳动保护科学 (安全科学)
	12 国际问题研究		Z 综合性图书

从表 2.1 可以看出, 《学科分类与代码》与其他三个分类表在一级学科的类目上、名称上等有許多相同之处, 但在一级学科的数量和类名上亦有一些不同。

(1) 一级学科的数量不同。《学科分类与代码》有关人文社科的一级学科是 20 个类。国家社会科学基金委的学科分类是 22 个 (出于评审等管理上的需要, 艺术学、教育学、军事学未纳入体系)。研究生的学科、专业目录则有 29 个一级学科, 该目录分成哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、军事学、管理学 12 个门类, 89 个一级学科, 386 个二级学科。其中, 人文社会科学 6 个门类, 加上属于综合门类的军事学、管理学则为 8 个门类, 29 个一级学科 (其中管理学 5 个), 二级学科 142 (其中管理学 14 个)。具体学科是: 哲学, 1 个一级学科, 8 个二级学科; 经济学, 2 个一级学科 (理论经济学、应用经济学), 16 个二级学科; 法学, 5 个一级学科 (政治学、社会学、民族学、马克思主义理论), 31 个二级学科; 教育学, 3 个一级学科 (教育学、心理学、体育学), 17 个二级学科; 文学, 4 个一级学科 (中国语言文学、外国语言文学、新闻传播学、艺术学), 29 个二级学科; 历史学, 1 个一级学科, 8 个二级学科; 军事学, 8 个一级学科, 19 个二级学科; 管理学, 5 个一级学科 (管理科学与工程、工商管理、农林经济管理、公共管理、图书馆、情报与档案管理), 14 个二级学科。被一个或两个分类表提到的学科有宗教学、考古学、民族学、人口学、心理学、统计学、管理学、党史·党建、国际问题研究、社会科学总论、环境科学、人文地理。

(2) 有的一级学科的名称不同。例如, 马克思主义类有四种名称: 马克思主义, 马列·科社, 马克思主义理论, 马克思主义、列宁主义、毛泽东思想、邓小平理论。另外, 民族学, 有的称为民族问题研究; 新闻学与传播学, 有的称为新闻传播学; 图书馆、情报与文献学, 有的称为图书馆、情报与档案管理; 管理学, 有的称为管理科学与工程。

(3) 有的一级学科排列的次序不同。如马克思主义类, 在国标和基金目录中位列第 1 类, 在研究生学位学科分类中排在法学门类下列第 8 类; 政治类、法学、历史学等的位次都不同。

(4) 有的一级学科分合不同。例如, 语言学, 有的合为语言、文字; 文学, 有的则分为中国文学、外国文学, 或中国语言文学、外国语言文学; 经济学, 有的分为理论经济、应用经济; 政治学, 有的则与法律合并为政治、法律; 历史学, 有的分为中国历史、世界历史, 有的合为历史、地理。

从各种学科体系的相同点可以看出, 我国人文社会科学学科体系已初步形成, 文史哲政经法等主要的学科已得到认同。例如, 中国当代权威的工具书《中国大百科全书》涉及人文社科的类目有《哲学》《宗教》《中国历史》《外国历史》《政治学》《法学》《军事》《经济学》《财政·税收·金融·价格》《社会学》《民族》《考古学》《文物·博物馆》《中国文学》《外国文学》《美术》《音乐·舞蹈》《戏剧》《戏曲·曲艺》《电影》《语言文字》《图书馆学·情报学·档案学》《教育》《体育》《新闻出版》《中国地理》《世界地理》等。分析各种学科体系的不同点, 可以看出现存的学科体系还存在学科名称、术语不够规范、学科数量不够齐全、学科之间的逻辑性需要加强等问题。随着科学技术的发展, 新学科和交叉学科不断涌现和高度综合, 而人文社会科学又具有民族性、历史性、区域性等特点, 如何将这些新情况和特点反映出来, 以进一步优化我国人文社会科学学科体系结构, 促进科学的繁荣发展, 是目前亟待解决的问题。要解决这些问题, 深入考察国外主要国家的学科体系, 无疑是必要的。

5. 国外主要国家人文社会科学学科体系分析

根据国务院学位委员会 21 次会议提出进行学科专业目录调整和调研国外主要国家教育学科体系情况, 上海交大、浙江大学、同济大学、哈工大、西安交大、延边大学等分工调研了美国、英国、德国、俄罗斯、韩国、日本的学科情况, 表 2.2 是根据网上资料, 经过笔者分析、增删、核查, 整理出的这些国家和联合国教科文组织“国际教育标准分类”(ISCED, International Standard Classification of Education) 文科及交叉学科设置情况的比较表。

表 2.2 各国（地区）学科门类设置情况比较表

学科门类	美 国	英 国	德 国	俄 罗 斯	韩 国	日 本	联合国 ISCED
人文学科 和艺术	英语语言文学	语言学·名著及相近科目	语言和文化科学（民族学、民俗学）	语文学/语言学/文化学	言语科学	文学语言/语言文化	外国语言与文化/语言学
		欧洲语言·文学及相近科目	跨学科研究（以语言和文化科学为主）				语言及文学
	外国语言文学	东方、亚洲、非洲、美洲、澳洲语言·文学等	文学				比较文学
	艺术学	创造艺术和设计	艺术，艺术学	艺术学/艺术	艺·体能	艺术	艺术学
	哲学与宗教	历史和哲学研究	哲学	哲学/宗教学	哲学	哲学	哲学/伦理学
	历史学	历史学	历史学	历史学	历史学	历史学	历史学
	神学	神学和宗教研究	基督/天主教神学	神学	宗教学/神学	宗教学科	宗教与神学
教育	教育学	教育	教育学/体育	教育学/体育	教育/体育	教育/体育	教育学
社会科学	社会科学（综合）	社会科学（综合）	法学、经济学和社会科学			社会科学	社会科学·商学和法学
	社会科学（其他）		社会科学				
	经济学	经济学	经济科学	经济学	经济学	经济学	经济学
	社会学	社会学	社会学	社会学	社会学	社会学	社会学
	政治学与政体	政治学	政治科学	政治学	政治、外交学	政治学	政治学/未来学
	考古学	考古学					考古学
	犯罪学						
	人口统计学与人口研究						人口学
	国际关系与国际事务				国际学科/国际关系		和平与冲突研究/人权
	人类学	人类学			人类学		人类学/民族学
	地理学与地图学	人文地理			地理学科		
	城市问题研究						
	工商管理学	商务和管理研究		管理学/商业	经营学科（MBA）		商业及管理学
	公共管理与社会服务		行政管理科学		行政学科		

续表

学科门类	美 国	英 国	德 国	俄 罗 斯	韩 国	日 本	联合国 ISCED
社会科学	法学与法律职业	法律	法学	法学	法学	法学	法学
	传播与新闻学	大众传媒和文献学		新闻学			新闻学和信息
	区域、种族、文化与性别研究		地区学	区域学/东方学、非洲学			区域研究、
	心理学	心理学（在生物学类下）	心理学	心理学	心理学		心理学
服务行业	家庭科学		营养科学和家庭学	服务	食品营养学/衣类学科	家政、生活科学（家政、食物、被服、居住）	家政学
	公园、娱乐、休闲、健身						
	个人与烹饪服务						
	安全与防护服务						
	交叉学科	综合学科		跨学科	学科间协同学科		
	文理综合						
其他学科	图书馆学	图书馆学	图书馆学、文献学与大众传播学	图书事业			图书馆学、博物馆、文献技术、档案科学
			社会福利事业	统计学	统计学		
				建筑学			

（注：上海交大等高校翻译整理的各国学科分类资料大多可信，但有的资料在译名、类目理解上有误，如英国的“L Social studies”被译为“社会学”，而 L Social Studies 类下又有“L300 Sociology”（社会学），还有“L100 Economics、L200 Politics、L300 Sociology、L400 Social Policy、L500 Social Work、L600 Anthropology、L700 Human and Social Geography、L900 Others in Social studies”，显然此处“L Social studies”应译为“社会科学”或“社会研究的各学科”，因为 Studies 是复数。又如，P Mass Communications and Documentation 中的“Documentation”被译为“文件”，从其下位类的类目也可判断，此处应译为“文献”^①。）

从表 2.2 可以看出，联合国及美、英、俄、德、日、韩设置的学科有以下特点。

（1）基本都有人文科学和社会科学。各国基本都有人文社会科学的主要学科：文史哲政法等。教育学、体育、新闻学、图书馆学、文献学大多数都有。

（2）设置交叉学科或跨学科或综合学科。美国、俄国集中设置交叉学科，而德国则在各大类下设置交叉学科或跨学科，例如，语言和文化科学类下设“跨学科研究（以语言和文化

① Joint Academic Coding System (JACS) v 1.7, JACS Complete Classification. http://www.hesa.ac.uk/dox/jacs/JACS_complete.pdf.

科学为主)”;法学、经济科学和社会科学类下设“跨学科研究(以法学、经济科学和社会科学为主)”。表2.3是美国CIP-2000学科专业交叉学科和综合学科类目表。

表2.3 美国CIP-2000学科专业交叉学科和类目表

学 科 门 类	序 号	下 位 学 科
交叉学科	1	生物与自然科学
	2	和平与对抗研究
	3	系统科学与理论
	4	数学与计算机科学
	5	生物心理学
	6	老年医学
	7	历史建筑、名胜保护
	8	中世纪与文艺复兴研究
	9	博物馆学
	10	科学、技术与社会
	11	会计学与计算机科学
	12	行为科学
	13	自然科学与生命科学
	14	营养科学
	15	国际/全球研究
	16	重大灾难与相关研究
	17	古老、古典、东方研究
	18	多文化与多元化研究
	19	神经科学
	20	认知科学
	21	交叉学科(其他)
文理综合	22	文理综合

在这些交叉学科中,有的是文理学科交叉,如会计学与计算机科学、系统科学与理论等。侧重于人文社科的交叉学科大多是专题的研究,如和平与对抗研究,国际/全球研究,重大灾难与相关研究,古老、古典、东方研究,多文化与多元化研究,等等。

(3) 设置服务、家政等应用性学科。如美国的家庭科学、德国的营养科学和家政学等。

(4) 文化学/文化科学、区域研究/区域学、人文地理、心理学、人类学、人口学、管理学/管理研究等学科被纳入人文社科门类下,而这些学科是否被纳入在我国还没有共识。

6. 构建中国人文社会科学学科体系框架的设想

建立既能与国际学科体系接轨,又能反映我国传统和现实的人文社会科学特点的学科体系是个复杂的系统工程,下面主要根据对国内外学科体系调研的情况,对我国现存的学科体系存在学科名称、术语不够规范、学科数量不够齐全、学科之间的逻辑性需要加强等问题提出改进建议,为构建我国人文社会科学学科体系的总体框架,即基本部类和大类提出设想。

1) 规范有关术语

作为与自然科学相对应的一个术语,目前在我国有几个术语:人文社会科学、哲学社会

科学、社会科学（广义）、文科等，这些术语经常混用，带来许多不便。尽管目前对人文社会科学这一术语的定义和所属学科没有完全统一，但基本的共识已经存在。根据我国国标学科分类中使用“人文社会科学”，国外也大多使用这一术语的情况，说明“人文社会科学”这一术语已经约定俗成，可以作为规范词使用，即在学术界均用这一术语，其他术语在不同场合也可以继续使用或逐渐废止。人文社会科学是以人类社会和人的精神为研究对象，探讨社会发展规律和人的生命本质的科学，由人文科学和社会科学组成。人文科学是以人类精神为研究对象，研究人的本质、价值的科学，包括哲学、文学、史学、语言学等。社会科学是以社会现象为研究对象，研究与阐述各种社会现象及其发展规律的科学，包括政治学、经济学、社会学等。其他大类的尚未规范的术语，如马克思主义，马列·科社，马克思主义理论，马克思主义、列宁主义、毛泽东思想、邓小平理论；民族学，民族问题研究；新闻学与传播学，新闻传播学；图书馆、情报与文献学，图书馆、情报与档案管理；管理学，管理科学与工程等。在同行专家、学者充分讨论后，选择几个术语中的一个或自选一个术语为规范术语，并对概念的内涵和外延给出明确的界定。

2) 调整类目结构

目前人文社会科学只有人文科学与社会科学两个部类，根据国外学科分类和学科发展情况，建议调整部类结构，由两个增加到五个，即综合学科（人文与社科的综合）、人文科学、社会科学、交叉学科（人文与社科下属各类的交叉）、应用学科。如联合国分成人文科学、社会科学、服务行业、其他四部类；美国分交叉学科、人文科学、社会科学、应用专业四部；日本则分为五大领域：人文科学、社会科学、家政、生活学、教育学、艺术学。我国经过30多年的改革开放，社会的经济、文化及有关学科都有很大发展，实用性社会科学发展也很快，并有很大潜力。应用学科可由从社会科学中剥离出的一些学科如教育学、新闻学等和我国正在新兴的一些实用、职业性专业或领域如家政学、服务学等组成。五大部类的排列可以从总体到具体，并反映出类与类之间的过渡性质，以增加学科体系的逻辑性。

3) 增设特色学科

中国有着深厚的文化底蕴，悠久的历史、睿智的思想、丰富多彩社会实践为人文社会科学学科建设提供了丰富的资源，也已产生了一些独特的学科。例如，中国古代的人文社会科学讲究整体化，文史哲不分，甚至经济学、政治学等社会科学的内容也糅在一起，这就是所谓的国学。目前国学又在大陆兴起，一些高校已设置专门的学院，甚至到海外合办有关学院，因此“新国学”作为综合学科可以增设到学科体系中来。又如，中国近几十年来发生了天翻地覆的变化，这与中国人研究、应用一门新的学科——马克思主义有密切关系，目前本土化的马克思主义已成为一门独特学科，也应在学科体系中反映出来。其他一些学科如文化学/文化科学、区域研究/区域学、人文地理、环境科学、心理学、人类学、人口学、管理学/管理研究等学科在我国人文社科体系中不明确，有的有，有的没有，但在国外体系中大多都有，因此建议这些学科也应被纳入进来。

诚然，对增设到学科体系中的学科要有一定的条件，这些条件包括要具有相当多的高校设有此学科/专业、有一批专业人员、有相应的学术组织、有专业期刊和研究成果等。只要具备这些条件，按照一定的审查程序就可以增设，这样可以将学科体系建设成一个动态的、开放的系统。

4) 完善学科体系框架

尽管我国国标《学科分类与代码》还有不完善之处，但毕竟有了一个学科体系的基础，在此基础上，加以修订，形成体系框架还是可行的。因为自1993年该标准颁布和执行以来，在全国已有相当大的影响，学位学科分类目录、社科基金目录、图书分类目录等都深受其影响。而且，该体系相对来说，简明扼要、结构清晰，其主要的学科大类已包括在内。建议新

的学科体系框架还可再概括一些,只列出基本部类和基本大类(一级学科)即可,最多分析到二级学科。学术共同体认可的新学科体系可以作为一个“总纲”,“总纲”相对稳定,根据学科发展变化情况定期修改。

5) 根据不同的目的与需求,构建不同的学科体系表

没有大体上共同认可的学科体系框架,各行其是,易造成混乱,但仅指望一个学科体系就能满足不同需求亦不现实和不必要。“和而不同”,各种学科分类在基本部类或大类设置基本一致,尔后针对不同的目的与需求,安排不同的类目。如学位学科目录主要体现培育人才的特点,文献学科分类则要反映文献组织和检索的需要,而科研学科分类目录则既要反映研究问题的学科性,又要反映以问题为导向的特点。目前国家社会科学基金委的学科分类表在这方面就有尝试:在一级类中,除按学科类列出哲学、政治学等大类外,也列出了“党史·党建”、“民族问题研究”、“国际问题研究”等主题/问题类目。这些主题/问题类目的设置主要考虑到了研究的需要。



本章小结

本章主要讨论了信息组织的理论基础和方法基础,其中,信息组织的理论基础包括有序化理论(包括自组织理论即新三论)、信息构建理论、知识组织理论、知识论、本体论等;信息组织的方法基础主要包括语言学、逻辑学和知识分类学。在信息组织的知识分类学方法基础部分,本章在分析现存国内各知识分类体系和国外人文社会科学学科体系的基础上,提出了构建中国人文社会科学学科体系框架的五点设想。



问题讨论

1. 有序化理论中与信息组织的关系最为密切的有哪些理论?
2. 什么是信息构建?请简述信息构建与信息组织之间的关系?
3. 本体论在网络信息组织中有哪些应用?
4. 概念之间存在着哪几种关系?概念的划分方法和规则是什么?
5. 亚里士多德、培根、恩格斯、毛泽东及钱学森的知识分类各有什么特点?
6. 形式逻辑与辩证逻辑在信息组织中各有何作用?



第3章

信息描述语言


本章引言

信息描述 (Information Description, Information Representation), 是网络时代学术界对传统“文献编目”概念的一种继承和发展, 是依据一定的规则 and 标准, 对存储于一定物理载体或网络上的信息的外部特征和内容特征进行选择、描述并予以记录的过程。信息描述语言是对信息外部特征和内容特征进行揭示和描述的语言工具, 是有效进行信息组织的语言工具。

按照所使用的语词标识是否经过人工规范, 传统信息描述语言可分为规范语言 (Controlled Languages) 和自然语言 (Natural Language) 两种类型。

分类法是一种按照类目之间关系组织起来并配有一定标记符号的分类信息资源的工具。主题法则是以表达主题内容的语词为检索标识, 以字顺为主要检索途径, 以参照系统等方法揭示词间关系的标引和检索信息资源的方法, 这两者各有所长, 相辅相成。

本章重点

- 规范语言的词汇控制方法;
 - 规范语言的类型与结构原理;
 - 自然语言标引的方式;
 - 分类法原理;
 - 分类法的标记系统;
 - 主题法的概念、原理及类型。
- 

3.1 信息描述语言概述

信息描述的目的是为了信息检索的需要，因此传统的信息描述语言也称为检索语言或标引语言。信息描述包括对信息外部特征的描述和对信息内容特征的标引，从而形成不同的检索标识以提供多途径的信息检索服务。

信息特征是信息所固有的、可借以确认某一或某些信息并将其从信息资源集合中识别出来的特征，包括信息外部特征和信息内容特征。信息外部特征的描述主要是将信息实体的题名、著者、出版者、出版时间、出处、报告号、专利号、网址等著录下来，并将不同的信息按照题名、著者名、出版者名称的字序进行排列，或按照出版时间、发布时间等时序进行排列，或按照报告号、专利号的数字进行排列，形成以题名、著者名、出版者名及号码等为检索标识的检索途径。信息内容特征标引是对信息所论述的主题、观点、见解、结论等进行揭示，并用专门的语词或符号表示，以提供内容检索途径。

在信息检索中，不管从何种途径入手进行检索，其最终目的都一样，即为了查找到具有特定内容的信息。内容检索途径是不可或缺的基本检索途径，而其他非内容的外部特征检索只是辅助的检索途径。为了解决信息内容标引工作的规范化、一致性问题，在图书情报工作实践中创制出了情报检索语言。情报检索语言是根据信息检索的需要而创制的人工语言，专门用于各种手工和计算机化的信息检索系统，表达信息的主题概念和用户检索需求的主题概念。情报检索语言是对信息内容进行描述的语言工具，是经过人工规范的，所以也称为规范的信息描述语言。

3.1.1 规范语言

1. 规范语言的定义及相关概念

规范语言是从自然语言出发，根据信息描述与信息检索的需要，从自然语言中筛选出特定的词汇来网罗和指示概念，并依据一定的规则对自然语言进行事先规范而形成的人工语言，是符号化的概念表示系统。因为它经过人为的规范和控制，所以规范语言也称受控语言（Controlled Languages）、人工语言（Artificial Languages）。

这些规范化的标识能较好地对表示概念的词汇或符号进行规范和控制。在信息管理领域，规范语言作为描述文献特征和表达用户信息检索需求的一种专门的人工语言，主要用来表示信息或需求的主题概念，为信息加工者和信息检索者之间提供一种共同语言，以便两者之间的交流和表示的一致，使用规范化的标识能相对有效地提高信息检索的效率。

例如“飞机”这一概念，在英文检索时，可用 Plane、Airplane、Aeroplane、Aircraft 等同义词来表示，但在规范语言系统中只能根据特定规则选定其中最适合的一个词汇来表示这一概念，如若选定 Aircraft 一词为规范词，则其余词均为非规范词。当用户使用“Aircraft”来检索时，其结果将包含所有有关“飞机”这一概念的文献，而不考虑这些文献中是否确切出现过“Aircraft”这个词。这样，用户在检索时可省略对其概念的全部同义词或近义词的考虑，也避免了这些词在输入时的麻烦和出错，规范语言提供了一种比较高效、能有效避免漏检、误检的检索途径。

规范语言是为了适应这种信息检索的需要而创制的一种人工语言，在手工检索条件下产生，并在机检环境下得到了充分的发展和应用。在数据库专业检索中普遍使用规范语言及其词表，凡是支持规范语言的检索工具，在内容检索时首选的是规范词检索。

规范语言因其主要用于对信息内容特征的揭示与标引,以建立信息主题索引系统,也称为标引语言或索引语言;其最终目的是满足信息检索之需,通常也称为检索语言或情报检索语言;在网络环境下各种传统信息描述语言和新型知识组织工具都被统称为知识组织系统(KOS, Knowledge Organization Systems)。

2. 规范语言的组成

规范语言在它的产生和发展过程中,以语言学、逻辑学、知识分类理论为基础,广泛吸收和引进了其他相关学科的研究方法。它与其他语言一样,也是由词汇和语法两大部分组成的。规范语言的词汇是指登录在分类表、叙词表、代码表中的全部标识,一个标识,如分类号、检索词、代码就是它的一个语词,而分类表、叙词表、代码表则是它的词典。规范语言的语法是指如何创造和运用这些标识来正确表达信息内容和信息需要,以有效地实现信息检索的一整套规则,分为词法和句法两部分。规范语言的词法主要用于分类表、叙词表、代码表编制过程,是创造和改革语词所使用的方法之和;句法主要用于信息标引和信息检索过程中,是将规范语言词表中的语词构成索引标识或检索标识的方法和手段的总和。

规范语言与自然语言不同,它根据信息检索的特定需求制定和设计语法,再构成词汇的集合。即规范语言通过制定词法规则来创造词汇或从自然语言中择取出可用的自然语言语词组成词汇集合,并建立词间的语义关系,再通过一定的句法规则来规定这些词汇在信息描述和信息检索过程中的使用。

规范语言是人工控制的语言系统,词汇控制和句法控制是其核心所在。

1) 规范语言的词汇控制

规范语言由语词标识构成,由于表述信息内容的词汇非常庞杂,日常使用的自然语言语词存在着一词多义、一义多词、词汇表达概念模糊和不确定、词汇量庞大、词间关系不明晰等多种问题。因此,必须通过规范化处理来实现对语词的控制,即词汇控制。所谓词汇控制,是根据信息描述和检索的需要,对自然语言中的词汇进行选择、规范并揭示其相关性的过程。通常的词汇控制包括词汇的选择、词形的控制、词义的控制和词间关系的控制。

(1) 词汇的选择。

鉴于规范语言词表容量有限,必须根据信息描述与检索的需求,对海量的自然语言语词进行精选和压缩,保留具有检索意义的词汇,以便尽可能地控制词汇规模。规范语言词汇选择应遵循以下几点原则:

① 应根据信息描述和信息检索的实际需要选词,尽可能选择检索中使用频率高并能汇集一定量文献的名词或名词性词组;同时,选词时应考虑学科或专业领域的现状及发展,对有关新兴学科或技术的词汇应考虑选用。

② 选用的语词应概念明确,一词一义,符合科学性、通用性的特点。

③ 词汇选择应考虑其组配性能,尽量收录核心词和具有构词功能的词,如选用“经济”而不用“经济学”;同时要注意适度原则,要求既能发挥组配的优越性,又能兼顾词汇的专指性,适当选用先组词,如“工业橡胶”用“工业”和“橡胶”组配会有“工业橡胶”和“橡胶工业”两层意思,故应考虑收录“工业橡胶”这一先组词。

(2) 词形的控制。

词形的控制主要是对词汇不同的书面表达形式及同义词或准同义词的控制,以实现每个语词的词义和词形的唯一性。

词汇不同书面表达方式的控制体现在以下几个方面:

① 规定词语形体,当一个语词存在几种形体,一律以通行的形体为标准,如用“储存”而不用“贮存”。

② 规定外来语译名的选择问题，一般采用通用译名而不用外来语的缩略语或音译词。

③ 规定数字和标点符号的用法，如选用“九·一八事变”，而不用“九一八事变”或“918事变”。

④ 规定词序，收录词表的复合词一般采用自然语序，不使用倒置形式，如用“工程物理学”而不用“物理学，工程”。

⑤ 规定外文的词形，规定外文的单复数、名词形式等，一般可数名词采用复数、不可数名词采用单数形式。

⑥ 规定词长，一般汉字词语控制在14个汉字以内。

为了能把论述同一主题概念的信息集中起来，需对同义词和准同义词实施控制。同义词的控制主要体现在学名与俗名、新名与旧名、全称与简称及不同译名上。而准同义词的控制则主要体现在部分近义词、反义词或宽泛词与专指词之间。

（3）词义的控制。

规范语言要求每个词汇只能表示一个概念，即词汇与概念一一对应，这就必须对多义词、同形异义词和词义含糊的词汇进行规范化处理，如采用加限定词或加注释的方法。

① 加限定词，一般用于自然语言中的同形异义词和多义词，通过从词汇使用范围的角度加以限定，明确词汇含义，使其具有单义性。如“病毒（医学）”、“病毒（计算机）”来明确“病毒”的含义。

② 加注释，对语词进行注释说明，包括含义注释、用法注释、历史注释等。

（4）词间关系的控制。

在整个规范语言词典中，无关联的语词只占极少数，大多数语词之间都是有关联的，都可以纳入到特定的范畴体系或族系中，建立与其他词汇之间的关系，形成一个语义关系网络。在规范语言中，通常会采用分类性质的索引、图示系统或参照系统来揭示词典中语词之间的关系。

规范语言的词汇控制体现在其编制过程中，根据控制程度的不同，形成了不同的规范语言系统。图3.1是根据受控程度的不同总结出的13种主要的规范语言，横坐标展现了规范语言受控的程度和受控的内容；纵坐标标识了规范的结构化程度，受控越强，结构化程度越高。

2) 规范语言的句法控制

句法是一整套组词造句的规则，信息描述语言中的句法特指用规范语词或符号来描述信息内容或检索需求主题时所使用的排序方式和规则。

在规范语言中引入句法控制有助于不同的人对同一信息内容的主题标引保持一致性；有助于用户掌握信息描述的规则以提高规范语言在信息内容表示上的适用性；有助于消除歧义，准确表达信息内容的主题。

规范语言句法控制的手段主要体现在引用次序、控制符号和句式变换上。

（1）引用次序。

引用次序（Citation Order）也称组配次序（Combination Order），是指用规范语言对复合主题进行标引或检索时，各个主题因素的组合或排列次序。目前主要的引用次序方案有：显著性引用次序、范畴职能引用次序和上下文从属引用次序。

显著性引用次序是按照各个主题成分的重要性或具体性递减确定的次序。美国学者克特（Charles A. Cutter）最早涉及显著性引用次序，在其1876年出版的《字典式目录规则》中有集中体现。克特之后，英国学者凯塞（J. Kaiser）对显著性引用次序做了进一步探讨，提出了更为具体的显著性次序。凯塞的思想被英国图书馆学家科茨（E. J. Coates）继承和发扬，建立了“事物—行为”的第一级显著性引用次序、“事物—材料—行为”的第二级显著性引

用次序和“事物—部件—材料—行为—施动者”的第三级引用次序。显著性引用次序在主题标引中具有很大的实用意义。

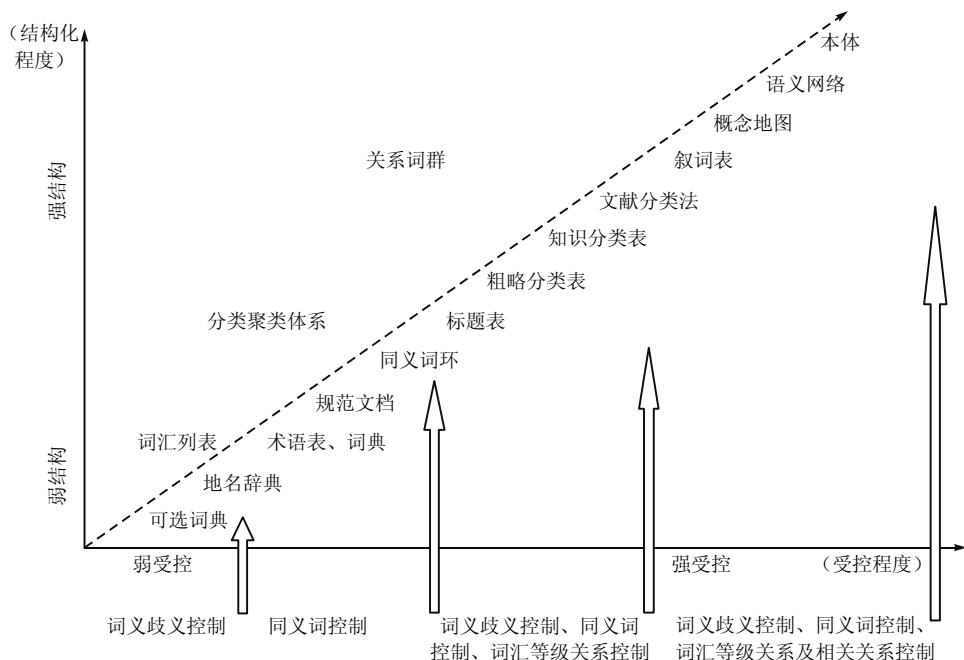


图 3.1 不同类型规范语言受控程度与结构关系图

范畴职能引用次序是用范畴这一思维形式认识主题，把各种主题概念划分为一系列的范畴，并根据范畴职能的具体性递减原则构成范畴职能引用次序。印度图书馆学家阮冈纳赞(S. R. Ranganathan)的范畴分面公式和英国的分类研究专家维克利(B. C. Vickery)的标准引用次序最具代表性。阮冈纳赞依据各学科领域的主题特征，将主题因素按性质概括为本体(Personality)、物质(Matter)、动力(Energy)、空间(Space)和时间(Time)五大基本范畴，并按照具体性递减原则来确定引用次序，形成 PMEST 分面公式。维克利在阮冈纳赞分面公式的基础上提出了更为具体的范畴划分，根据学科特点的不同划分不同的学科范畴，并设计了“物质(产品)—结构—成分—性质—材料—过程—空间—施动者或工具”这一引用次序。

上下文从属引用次序是伴随着索引编制自动化的发展而逐步建立起来的一种接近自然语言句法和语序的引用次序。其中以英国学者奥斯汀(D. Austin)研制的保留上下文索引(PRECIS, PREserved Context Indexing System)最具代表性。PRECIS 放弃将最显著的语词置于款目首位的传统做法，而是根据上下文从属原则拟定标引语句。上下文从属原则是根据复合主题各个主题因素的关系，按照从宽到窄的次序予以排列，即在整个标引语句中，前一个术语将后一个术语置于比较广阔的语境中，使其按易于理解的次序加以组织。

我国著名情报语言学家张琪玉对文献主题的构成因素和层次结构进行了研究，分类学家刘湘生则提出了中文文献的主题分面公式：

A 主体因素(A1 研究对象、A2 材料、A3 方法、A4 过程、A5 条件)—B 通用因素—C 空间因素—D 时间因素—E 文献类型因素。

这一主题分面公式已作为中文信息主题标引引用次序的国家标准，指导中文信息主题标

引中的句法控制。

（2）控制符号。

在自然语言的语句中，人们常用一些功能词，如介词或连词对句法结构起控制作用。而在信息描述的规范语言中，句法控制不是靠这样一些功能词，而是通过采用人工规范的一套控制符号来实现的。控制符号主要有联系符号、职能符号、关系符号和加权等。

① 联系符号，又称联号，是解决虚假组合的控制符号。即在标引某一信息时，论及多个主题或某一主题被赋予多个标引语词时，应用联号来标识标引语词之间的组合关系。如“电脑组装与电视维修”这一并列主题，如果简单地用“电脑”、“电视”、“组装”、“维修”标引时，那么就有可能产生虚假组合：“电脑—维修”和“电视—组装”，而使用联号则可以避免上述错误组配。以数字 1、2 为联号，那么“电脑组装与电视维修”这一文献信息（文献号为 0001）可标引为如表 3.1 所示的形式。

表 3.1 联号在信息主题标引中的使用示例

主 题 标 识	联 号	文 献 号
电脑	1	0001
组装	1	0001
电视	2	0001
维修	2	0001

联号最初广泛使用在单元词检索系统中，美国杜邦公司的单元词卡片系统就是用一个较简单的联号进行句法控制的。

② 职能符号，简称职号，是一种表示主题标识在组配中的句法职能的辅助符号，用来显示组合过程中每一个主题标识在句法结构中的句法意义，以便精确组配。例如，某文献论述“气候影响森林”，另一篇文献论述“森林影响气候”，这两篇文献都可以用“气候”、“森林”和“影响”三个语词标引，但实际上在这两篇文献中这三个语词的关系意义是不一样的，须使用职号来区分。根据句法结构需要，设定职号：a 代表受动者，b 代表操作过程，c 代表施动者，则：

气候影响森林 标引为 气候—c； 影响—b； 森林—a

森林影响气候 标引为 森林—c； 影响—b； 气候—a

③ 关系符号，是通过对信息主题标引语词两两间存在的关系的揭示来控制的符号。如在《国际十进分类法》中的并列符号“+”揭示了两个主题的联合关系，如用“54+56”表示“化学和化工”，用“537+621.3”表示“电学与电工”这样的复合主题之间的关系。

④ 加权，是对表示信息内容的语词赋予加权符号或权值来控制其句法意义和功能的控制方法。加权可用于标引也可用于检索，加权标引是为每个标引词的重要性赋予不同的加权符号或权值；加权检索是要求检索出的信息其词权之和能达到一定的阈值，否则不予命中。

（3）句式变换。

在手工检索时期，检索表达式中句首词起着引导检索者的作用，而句中其他的词因为不在句首而不能起到这个作用，因此需要用到句式变换，对其他有检索意义的词进行轮排，将其作为入口词放到句首以充当检索入口。到了机检时代，支持全文检索时，这种句法控制手段的意义就变得无足轻重了。

3. 规范语言的功能

规范语言产生于手工检索阶段，在机检时代得以发扬，在目前的网络检索中仍占有一席

之地。规范语言在信息检索中起着极其重要的作用,它是沟通信息描述与信息检索的桥梁。在信息描述过程中,规范语言用来表示信息的内容特征和部分外部特征,从而形成检索标识;在信息检索过程中,规范语言用来描述和表达检索提问,从而形成提问标识;当检索标识与提问标识匹配时,结果即为命中。规范语言作用于信息的描述、组织和检索的整个过程中,起着揭示信息内容和组织信息检索系统的重要作用。

1) 标引功能

对信息内容特征和部分外部特征的描述和标引是规范语言最基本的功能。采用规范标识,如分类号、主题词来表达文献主题概念,使文献主题概念的表达规范化,能保证不同标引人员描述信息的一致性。

2) 揭示功能

规范语言的这一功能体现在其聚类功能上,通过一定形式揭示规范语言所表达的文献主题之间的相同性、相似性和相关性,以构成一个文献主题概念网络,把知识和信息纳入这个概念网络中,达到“物以类聚”、“鸟瞰全貌”和“触类旁通”的作用,有助于文献标引人员准确选择标引用语,从而提高标引质量;也有助于信息检索人员准确选择检索用语进行族性检索,并能根据检索的具体情况进行扩检、缩检,从而提高检索效率。

3) 整序功能

利用规范语言的系统性将文献信息的款目按照检索标识进行集中化、系统化和组织化,便于检索者按照一定的排列顺序进行有序化检索,从而提高检索速度。

4) 比较功能

通过规范语言的语词标识和句法控制,既能保证不同加工人员描述信息的一致性,又能保证检索者与加工者对相同文献内容表述的一致性。

4. 规范语言的类型

1) 按结构原理分

规范语言按照其结构原理不同可分为分类语言、主题语言和代码语言三种基本类型。

(1) 分类语言。

分类语言用分类号来表达各种概念,将各种概念按学科属性进行分类和系统排列,是一种按照学科范畴划分而构成的语言体系。分类语言集中反映了学科的系统性,反映了概念的相关、从属、派生等关系,按照从总到分、逐层分面展开,形成层级式分类体系。文献分类法、商品分类表、专利分类表等都是常用的分类语言。分类语言包括等级体系式分类语言(Hierarchical and Enumerative Classification; Systematic Classification)、分面组配式分类语言(Facet Classification)和等级—组配式分类语言。

(2) 主题语言。

主题语言用语词来表达各种概念,并揭示语词或概念之间的等同、等级、相关等关系。主题语言包括标题型主题语言(标题法)、单元词型主题语言(单元词法)、叙词型主题语言(叙词法)和关键词型主题语言(关键词法)等。其中关键词主题语言属于非规范化的主题描述语言,即自然语言。

(3) 代码语言。

代码语言是指对事物的某方面特征,用某种代码系统来表示和排列,从而提供检索的信息描述,如化合物分子式、档案号、专利号等。根据化合物的分子式这种代码语言,可以构成分子式索引系统,允许用户从分子式出发,检索相应的化合物及相关的文献信息。

引证关系追溯法是显示科学文献之间相互引证而形成的一种网状关系,通过文献信息之间的引用与被引用关系来揭示文献信息之间的主题关联。随着《科学引文索引》(SCI, Science

Citation Index）和 Google 的 PageRank 算法的推广，人们也把这种引证关系追溯法看做是一种信息描述语言的特殊类型。

2) 按其标识的组合使用法分

规范语言按其标识的组合使用方法，可分为先组式语言和后组式语言。

(1) 先组式 (Pre-coordination) 语言。

先组式语言是指语词标识在编表时就固定组配好，用户只能用这种已固定好的语词词组形式去描述信息内容或信息需求的一种规范语言。它有较好的直接性和专指性，但灵活度差，如标题词语言。

(2) 后组式 (Post-coordination) 语言。

后组式语言是指在检索实施前未事先组配好的一种检索语言，在检索时将它们临时组配起来，表达一定的概念来完成检索。这种后组方式提供了灵活的组配方式，在计算机检索中得到了广泛应用。

3) 按学科、专业范围分

规范语言按其所覆盖的学科或专业范围可划分为综合性语言和专业性语言；按适用范围可划分为国际通用的语言、各类图书馆和情报机构通用的语言、某一类型图书馆和情报机构专用的语言、某一类型信息资源专用的语言、某一检索系统适用的专用语言。

4) 按其他标准划分

除以上述划分方式外，规范语言还可以按其标识的语言文字类型划分为单语种语言和多语种语言；按其适用的环境分为传统环境下的信息描述语言和网络环境下的信息描述语言。

从上述划分看，由于对规范语言认识的侧重点不同，划分的标准也不相同，得出的结果不尽相同。

规范语言是为了克服自然语言的不足而产生的，在手工检索和机械检索环境下，对提高信息检索的查全率和查准率发挥了重要作用。但随着信息数量的激增，传统的规范语言标引难度大、速度慢、词汇更新滞后、对标引和检索人员要求过高等弊端日渐明显。随着计算机检索的发展，广大的网络终端用户对检索系统易用性、便捷性的要求，决定了自然语言能在广大潜在用户和网络用户中受到欢迎，在网络世界中拥有广阔的发展前景。

3.1.2 自然语言

1. 自然语言概述

无论是哪种规范语言，为了达到简明、专指地进行标引及准确、便捷地实现检索的目的，都会对人们日常使用的自然语言进行种种转换及限制处理，以达到控制的目的。如分类语言是建立在代表主题概念的一系列类目基础上的号码体系，而主题语言则是有选择并加以规范化的自然语言的一个子集。这些规范语言在克服自然语言的两大不足——概念与语词非一一对应及概念关系的隐含性的同时，也不可避免地导致了规范语言自身的局限，即表达概念的受限、词汇转换的失真、标引与检索前处理量大且难以达到统一等。

20 世纪 70、80 年代，由于计算机软硬件条件的支持，国外在联机检索的基础上进一步向网络化发展。20 世纪 90 年代以来，各国信息高速公路建设更是方兴未艾，如火如荼，国内机检水平也在迅速提高，并与全球大趋势相合拍。在这种形势下，规范语言因其固有的人工性、受控性，越来越难以满足大量、迅捷、自由、多样的检索要求，其得不偿失的内在不足更趋明显化，这时对自然语言的重视又开始日趋回温。信息的海量增长、用户的非专业化及计算机的高效处理最终引导了自然语言的回归。

1) 自然语言的概念

“自然语言”(Natural Language)是人类在社会生活中发展起来的用来交流的声音符号系统。从信息检索的角度来理解,自然语言是在信息描述和信息检索中可直接使用的、人们在科学交流中采用的书面语言或口头语言。自然语言包括关键词、自由词和出现在文章题名、摘要、正文或参考文献中的具有实质意义的词语。自然语言是非规范语言,直接采用未经人工控制的词语或符号作为标识。

自然语言语词具有较大的灵活性、专指性强,它能及时地反映最新出现的词汇,反映规范词难于表达的特定概念或新概念。在计算机全文检索中自然语言独领风骚,大容量、高速、高性能的计算机检索系统的自动标引,使得基于关键词的全文检索(Free-text Search)占计算机信息检索的比例越来越高。

2) 自然语言的复兴

自然语言作为日常使用的语言,在“以用户为中心”的网络时代无疑最符合人们信息检索的习惯和要求。随着机检的高度发展、应用条件的日益完善,自然语言有可能扬长避短,重新发挥其固有的优势。

在信息描述计算机化和信息检索用户非专业化的信息环境下,规范语言之于信息处理、信息系统和信息用户都存在诸多弊端。

由于计算机容量大、运行速度快、检索功能强等特点,在信息急速增长的今天,特别是大量涌现的网上信息,再也无法按部就班地使用规范语言进行信息处理,传统的规范语言受到了严峻的挑战;对数量巨大而复杂的网上动态信息,用规范语言进行信息处理显然是困难的,只能依赖于自然语言并借助计算机进行信息处理。

基于规范语言的信息检索系统,其信息描述和信息检索都依赖于特定的分类表或词表,首先对于使用者存在一个熟悉规范语言使用的认知负担问题,其次,对于基于不同规范语言描述的系统还存在一个跨库检索的兼容性问题。

互联网环境下,用户的信息需求往往不再诉求于专业的信息工作人员,而多半由用户自己来完成检索工作,易用、快捷成为人们对于信息检索系统最主要的要求。规范语言的严格规范性引起普通用户对于这种检索的排斥,尽管它的效率或许优于自然语言检索,但仍难以作为普通用户,尤其是互联网用户所接受。

基于这样一些因素,自然语言的直观性与专指性好、检索途径多、便于计算机自动抽词、标引速度快、容易掌握、检索方便、查准率高,适应科学技术发展等优点得到了信息用户的认可,自然语言开始在信息组织与信息检索领域复兴。

因此,自20世纪50、60年代进行的二次加菲尔德(Eugene Garfield)实验得出最少实施控制的系统较其他系统优越的结果后,自然语言检索系统得到了迅速发展,发达国家的联机检索已从只能利用规范语言进行布尔逻辑检索的第一代发展成能利用自然语言进行语境逻辑检索的第二代,而网络检索中自然语言已无可争议地成为主流。

3) 自然语言区别于规范语言的特点

自然语言在计算机检索和网络检索环境下开始复兴,促使自然语言复兴的原因除了计算机应用和网络的日益普及外,还归根于其自身的一些突出优点。

(1) 直接采用文献信息作者使用的自然语言,信息标引工作就可以摆脱规范语言繁复的分析转换过程,降低标引负担和成本,提高标引速度。

(2) 采用自然语言标引与检索,可达到足够的专指度,且不存在类目或词汇更新滞后的问题。

(3) 直接以日常使用的自然语言进行信息检索,符合检索者的习惯,简便易行,对于日

益增大的普遍用户群而言更是如此。

(4) 自然语言具有通用性，不存在规范语言的统一兼容问题，在使用自然语言的各项数据库间可实现标引、检索成果的共享。

(5) 自然语言标引为计算机的自动处理创造了条件，其发展将可能替代费时、费力的人工标引。

当然，自然语言也非十全十美的信息描述语言，张琪玉教授指出，自然语言在信息描述与信息检索的应用中面临着两大难题：一是如何从自然语言文本中抽取最能准确、充分地表达文献有价值内容的词，以及这些词与检索需求有效匹配的问题；二是如何克服自然语言由于不规范和缺乏语义关联性而对检索不利的问题。针对这两大难题，一方面要依靠计算机技术和自然语言处理技术的突破，另一方面，不论计算机技术和自然语言系统如何发展，受控语言的基本原理——词汇控制是永远不会消失的，变化的只是词汇控制的方式、方法和手段。

2. 自然语言的应用

1) 自然语言处理

自然语言处理（NLP, Natural Language Processing）是自然语言得以应用所要解决的首要问题。自然语言处理是人工智能领域的一个重要分支，主要研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法，是一门融语言学、计算机科学、数学于一体的科学。

自然语言处理要实现人机间自然语言通信，关键是要让计算机“理解”自然语言，所以自然语言处理也叫做自然语言理解。一般认为，自然语言处理主要有以下四个主要应用领域：机器翻译、信息检索、人机交互和篇章理解。这四方面的技术构成了自然语言处理研究内容的应用技术部分。本节将重点介绍自然语言在信息组织和信息检索中的应用，自然语言在信息组织和信息检索中的应用主要表现为自然语言标引和自然语言检索两个方面。

2) 自然语言标引

如前所述，自然语言较之规范语言更适于当前的信息状况及机检条件，并且数十年来在信息组织与信息检索实践中得到了相当的应用和发展。按照兰开斯特的划分，自然语言标引可分为自由标引、自动标引、不标引或全标引三种形式。

(1) 自由标引。

自由标引即人工关键词标引，由标引人员在对信息内容进行主题分析的基础上，按照一定的规则自拟标引词来表达信息主题。标引人员自拟的标引词属于自然语言，是一种不依据词表的主题标引方法。

自由标引克服了人工受控标引速度慢、周期长的缺点，降低了标引成本；自由标引不使用词表，所以可采用与信息主题专指度一致的词进行标引，能保证较高的查准率；自由标引由人工完成，如果标引人员具有一定的业务水平，那么其标引质量将大大高于自动标引。

“自由标引主要适用于报纸、期刊文献篇名数据库的标引，因为这类文献内容庞杂、新主题多、数量巨大，很难编制适用的词表，而且使用词表标引负担重、速度慢，建库单位实际条件往往不许可。”需要指出的是，自由标引虽然是一种自然语言标引方式，但在整个标引过程中并未涉及自然语言处理技术。

(2) 自动标引。

自动标引是指利用计算机从各种文献中自动抽取相关标识的过程。自动标引包括主题自动标引和分类自动标引两种，主题自动标引按标引词来源的不同，分为自动抽词标引和自动赋词标引；分类自动标引按照分类方式的不同分为自动赋号标引和自动聚类。图 3.2 揭示了各种自动标引方式之间的关联。

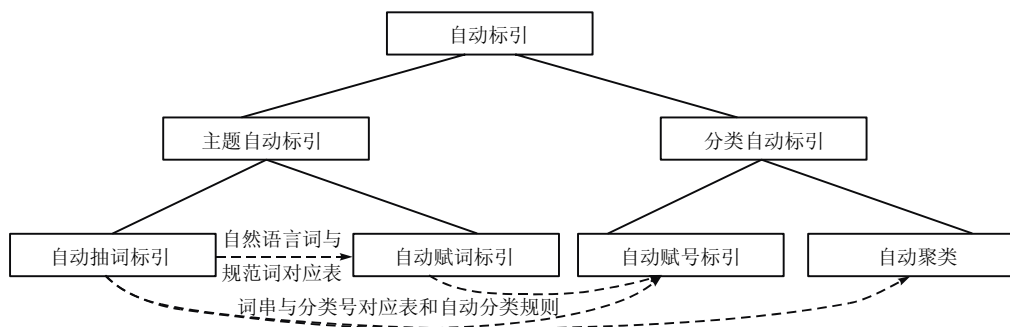


图 3.2 各自动标引方式之间的关联

① 自动抽词标引。自动抽词标引是指用计算机自动抽取信息资源中表达信息主题的语词作为检索标识，是最早出现的一种自动标引方案，由美国 IBM 公司的卢恩（H.P. Luhn）等人率先提出。自动抽词标引主要是通过与禁用词表的比对，从文献中自动抽取能表征文献主题的关键性语词作为标引词。汉语信息的自动抽词标引主要问题是汉语分词问题。

② 自动赋词标引。自动赋词标引是在自动抽词标引的基础之上，引入预先编制好的规范词表来规范从文献中自动抽取出来的语词，利用计算机的自动转换功能，将关键词转换为规范词，并赋予文献主题。赋词标引中的语词来自受控词表，所以从本质上讲，自动赋词标引是自然语言与规范语言相结合的一种标引方式。

③ 自动赋号标引。自动赋号标引即自动归类，是在自动抽词的基础上，根据自然语言语词与分类号的概念映射关系和相应的分类规则，利用计算机的自动转换功能，将关键词转换成分类号。因其最终检索标识来自分类表，所以从实质上讲，自动赋号标引也是自然语言与规范语言的结合。

④ 自动聚类。自动聚类是自动分类的另一种形式，是指由计算机考察待分类文献信息的内部或者外部特征，按照一定要求将这些特征进行计算比较，将相近、相似或者相同特征的信息对象聚合在一起的过程，它与自动赋号标引的不同之处在于它的类目体系不是预先编制好的，而是根据信息集合的内容聚合出来的。自动聚类建立在自动抽词的基础上，引入了自然语言处理技术，是自然语言在信息组织中的一个应用。

（3）不标引或全标引。所谓不标引方式，即对信息不进行任何标引，在检索时，借助计算机的自动匹配功能直接以关键字、关键词或词组作为检索用语，在文献标题、文摘或全文中进行匹配查找。如果数据库中存储的是全文信息，并且在全文中进行查找，则特称全文检索，这是自然语言检索最常用的方式。

与这种不标引方式相对应的另一种典型的自然语言标引方式称为“全标引”，是为了解决汉语分词问题而提出的单汉字索引。单汉字索引以单个汉字作为标引或检索单元，对文本中的每一个汉字都建立倒排索引，检索时用单汉字组配法查找。因为单个汉字在绝大多数情况下是不能独立表达文献主题内容的，因此这种“全标引”方式实际上等同于不标引。这种方式比较完全地实现了自动化，标引深度大、检索方便灵活，解决了汉语分词难题；但由于以字为处理对象，容易产生虚假组配，检索噪声大，筛选负担重，检索效率不高。这种单汉字索引方式不是真正意义上的标引，只是一种索引方式，在搜索引擎的全文索引中应用较为普遍。

基于上述的自然语言标引，相应的自然语言信息检索系统应提供自然语言检索接口，供用户输入自然语言进行检索。

3) 自然语言检索

自然语言检索是用户用自然语言作为提问输入到基于自然语言处理技术的信息检索系统的一种检索方式。

从用户输入检索词的形式区分,自然语言检索可分为关键词检索和自然语言语句提问式检索。关键词检索是用户提交其认为能表达其信息需求的关键性语词,由计算机在经过自然语言标引或全文索引的系统中进行匹配并返回相应的结果。而自然语言语句提问式检索则允许用户直接以日常用语的自然语句形式向系统提问,由带有一定人工智能的自然语言接口对这种提问式语句进行分析,然后返回相应结果。关键词检索是目前最常用的自然语言检索方式,尤其是在搜索引擎的检索中;而自然语言语句提问式检索则是自然语言检索系统所追求的目标,是真正实现自然语言理解的信息检索。

从检索的内容来看,自然语言检索分为基于自然语言标引词的检索和全文检索。基于自然语言标引词的检索是对以自然语言标引的内容进行检索,如文献篇名、作者、关键词等,它是对自然语言标引结果的检索,返回的结果可能是文献题录也可能提供全文;全文检索是对文本全文的检索,是一种建立在全文索引基础上的检索技术,这种检索技术无须对文献进行标引,它面向全文进行查找匹配,返回结果也是全文信息。

自然语言检索已成为网络信息检索的主流技术,现在越来越多的信息检索系统支持自然语言检索,但自然语言固有的弊端仍是影响检索系统效率的主要因素,建立真正基于自然语言理解的智能信息系统仍有许多需要解决的技术难题。目前,自然语言检索存在的不足主要表现在以下三个方面。

(1) 由于自然语言语词与概念不是一一对应的关系,无法排除同义词、近义词、多义词等对信息描述和检索的影响,在自然语言标引中,缺乏有效控制,会导致同一主题信息的分散标引,难以实现将内容相同或相近的信息加以集中并揭示其相关性的作用,影响查全率。

(2) 由于自然语言语词缺乏对词间关系的揭示与描述,自然语言检索时容易造成不相关结果的大量输出,检索噪声大,检索效率低下,影响查准率。

(3) 自然语言检索系统的检索效率直接与用户的认知水平相关,为了提高检索效率,用户认知负担加重,影响了系统的易用性。

所以,在这种情况下,有必要引入一种机制(即后控制机制)来弥补自然语言的不足。

3. 后控制词表对自然语言不足的弥补

在计算机检索越来越普及的情况下,自然语言具有不可阻挡的发展态势,特别是在网络检索环境中,它是一种必然的优先选择;另一方面,自然语言和规范语言并不是绝对对立,它们各有长处和短处,可以互相结合,互为补充。这对我们正确认识自然语言检索存在的不足、探寻有效的改进方法有很大的启发。

以往的规范语言在文献或信息描述时就对标引词先行加以控制,因此这种规范语言也称前控到词表(Pre-controlled Vocabulary)。这种控制带有一定的粗泛性、滞后性,有时甚至失控;因而人们开始尝试使用自然语言进行检索,但又造成了与规范语言的脱节。于是,在标引时使用自然语言,在检索时使用自然语言并实施一些不严格的控制,这就成了后控制词表(Post-controlled Vocabulary)的最初思路。

1) 后控制与后控制词表的概念

从整个信息检索的发展来看,控制是必要的。若要达到较高的检索效率,必须实施控制。

信息检索中的控制体现在信息标引和信息检索两个阶段。按照控制与否及控制所处阶段,信息检索的控制可分为四种模式。

(1) “标引控制+检索控制”,这种模式在信息描述和信息检索阶段都要采用规范语言来

进行标引和检索，是一种纯人工语言模式。

(2) “标引不控制+检索不控制”模式，这种模式在信息描述阶段和检索阶段都使用自然语言，不使用任何词表，是一种纯自然语言模式。

(3) “标引控制+检索不控制”模式，该模式在信息描述时采用规范的检索语言，检索时既可以使用规范语言又可以使用自然语言，在系统内部存储有一部自然语言→规范语言的转换词典，把检索用户的自然语言转化为受控语言，相当于在检索端增加了一个自然语言接口。

(4) “标引不控制+检索后控制”模式，这种模式下的标引阶段使用自然语言，检索时则既可以使用自然语言又可以使用存放在机内只供检索用的词表，即后控制词表来加以控制。

上述第一种和第三种模式属于先控制模式，第二种是非控模式，而第四种模式就是称为“后控制”的模式，属于自然语言检索法，所使用的词表称为后控制词表，是为了弥补自然语言不足而提出的一种控制方法。所以，后控制是一种在标引阶段使用自然语言，不对标引进行严格控制，而在检索阶段才对检索词进行控制的自然语言检索优化技术。

具有后控制词表意义的概念最早出现在 20 世纪 50 年代初的美国，后控制词表只用于检索而不用标引，当时也称为“只检索用词表”。后控制词表是利用规范语言的基本原理和方法编制的自然语言检索用词表，它主要对自然语言中大量存在的等同关系、等级关系和大部分的相关关系进行控制或揭示，它可以根据检索需要将新概念和新术语及时地加入到词表中去，是一个不断增长的“自然语言叙词表”。

2) 后控制词表的控制机理

后控制词表利用规范语言的受控原理来弥补自然语言检索的不足，是提高自然语言检索效率的有效措施。后控制词表在自然语言检索系统中的应用主要有三种方式：其一，用户检索前通过浏览后控制词表选择合适的检索词构造检索式，检索负担相对较重；其二，由系统自动执行调整检索式，既能减轻用户负担又能提高检索效率，但受限于目前的自然语言处理技术，效果不甚明显；其三，是前两种方式的融合，系统根据用户输入的自然语言检索式从后控制词表中给出相关词，供用户选择来调整检索式，通过这种交互式的相关反馈提高系统检索效率，是目前最常用的一种后控制方式。

后控制词表具有自然语言和规范语言的双重性质，它实际上是一种转换工具，在自然语言检索过程中对用户的自然语言检索要求加以分析、综合、归纳，转化为系统可接受的语言。后控制词表由控制词和被控制词组成，控制词是规范语言，但不是用来标引或检索的，而是用来对自然语言检索标识进行控制的；被控制词是自然语言，是用户在标引和检索中给出的自然语言标识，这与一般规范语言中的情形正好相反。控制词与被控制词的关系包括等同、等级、相关三种，其中等同关系的词（包括可视为等同关系的词）占大部分，这也表明后控制词表主要是为了解决自然语言语义含糊、同义、多义等问题。在使用后控制词表的检索过程中，用户或系统可以从任何一个词出发，在词表中查到该词的一批同义词、等级词、相关词。例如，用户想要查找有关儿童方面的文献，只要输入“儿童”一词，词表自动列出“少年”、“幼儿”、“男孩”、“女孩”等词汇，并用这些词语去辅助检索，这样用户就不必自己去设想所要查的主题到底有哪些同义词、等级词和相关词了，并且词表所提供的词的数量往往比较多，检索结果较好。图 3.3 揭示了后控制词表在自然语言检索系统中的应用。

众多实验表明，后控制词表不仅具有自然语言系统的全部优点，而且还具备许多先控规范语言的特性，是“自然语言检索和人工受控语言相结合的最佳范例”。

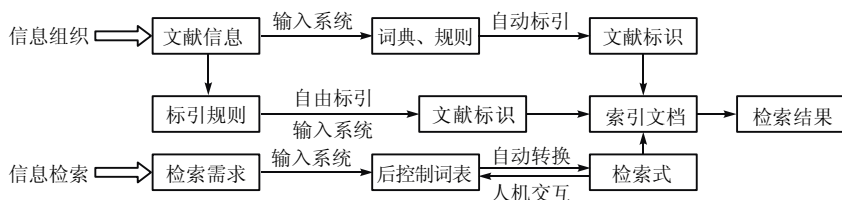


图 3.3 后控制词表在自然语言检索系统中的应用

3) 后控制词表的编制

后控制词表中的控制词并不直接用于标引,而是对作为文献检索标识的自然语言词语进行控制。因此,后控制词表必须在检索系统中实有的自然语言检索标识的基础上进行编制,即必须以作为检索标识的自然语言原词为基础,否则将会大大降低其控制功能。这一点是后控制词表编制的关键所在。

国内在这方面的研究始于 20 世纪 80 年代初,近年来伴随着理论研究和实践的深入,出现了一批实验性或实用性后控制词表系统。如基于用户提问和用户建议的自学后控制系统、基于字面相似原理的后控制词表辅助编制系统、基于词对相似和词对共现原理的后控制词表系统及基于分面分析的机辅后控制词表系统等。从这些研究不难看出,机辅或自动编表是后控制词表编制研究的主要方向,手工编表已不可行。

后控制词表的机器编制方式归纳起来主要有词典实现模式、积累提问式模式、词形实现模式、聚类控制实现模式、词频统计模式和人工智能模式等。

(1) 词典实现模式。

即利用某种现成的词表或分类表作为框架,把自然语言检索标识作为参照系统的“用”项纳入其中。

(2) 积累提问式模式。

利用计算机自动收集检索表达式中的用词并加以积累,然后由人工判别整理成词表。用这种方式编的后控制词表,覆盖率很低,要积累很长时间才能达到较高的覆盖率。

(3) 词形实现模式。

利用词汇的字面成族原理,将词形相似的词语集中,这在搜索引擎的相关反馈中应用得比较多,是一种利用计算机较易实现的后控制方式,但效果不佳,对于那种词形不同的同义词无法控制。如“番茄”和“西红柿”,词形完全不同,词义却是一样的,利用这种基于词形相似的方式,无法实现对两者的控制。

(4) 聚类控制实现模式。

采用聚类分析方法为自然语言检索系统中的词汇建立词间关系,形成语义网,以类别来引导一部分相同、相近或相关的词。

(5) 词频统计模式。

这种模式基于词汇在文献或检索需求中的共现统计来建立词汇间的关联,通过统计某一词对在文献中的共现频率,计算其关联度,然后进行阈值判断,将关联度高的词对提取出来进行分析,对词对之间可能隐含的关系进行合理的解释和预测。

总之,后控制词表是面向信息用户设计的,其编制应建立在检索系统中用户输入的实际自然语言检索标识的基础之上,以达到最大覆盖率,这是一个动态的累积过程。后控制词表被认为是目前所知的受控语言和自然语言结合的典范,它兼有传统的受控语言和自然语言的长处,是一种有发展前途的语言系统。

3.2 分类语言

3.2.1 分类法的原理

1. 信息资源分类

信息资源分类,是指根据信息资源的内容属性和其他特征,将其进行分门别类地、系统地组织和揭示的方法。一般说来,信息资源分类是以知识分类或学科分类为基础,结合信息资源各种载体的实际编制的类目体系,它与知识分类既有相同点,又有所不同。

要有效地对信息资源进行分类,就需要了解信息资源的本质属性与非本质属性。信息资源的本质属性是指载体上所记录的科学知识内容,它从根本上体现信息资源的价值、使用价值;信息资源的非本质属性是指与本质属性相对应的属性,一般体现在信息资源的形式特征上,如载体、语种等。

信息资源的分类是一种从主题内容角度组织和揭示信息资源的方法,是分类方法在信息资源组织中的应用。信息资源分类具有以下特征。

(1) 按照内容特征的相互关系对信息资源进行组织。

分类是指以事物的本质属性或其他显著特征为根据,把各种事物集合成类的过程。信息资源分类是分类方法在信息组织中的应用,是按照分类对象关系进行揭示的。不仅根据信息资源的内容属性进行区分和类聚,同时还将各种门类的信息资源按照类目之间的关系组织成一个具有等级性、次第性的系统。在这一系统中,各种类目根据等级次第进行排列,反映出信息资源的远近亲疏关系。用户可以根据特定的检索需求,按照知识之间的内在关系,有针对性地查找具有特定内容属性特征的信息资源,不仅如此,还可以依据类目之间的联系,灵活地扩大或缩小查找范围,并且可以实现对相关资料进行系统检索的目标。

(2) 从一定角度出发组织和揭示信息资源。

信息资源主题内容之间的联系往往不是单一的,而是多方面的、多维的。分类法是从内容角度揭示信息资源的方法,通常是有选择地揭示主题内容之间的主要联系。为了适合用户从学科的角度查找文献的习惯,目前的学术性文献分类体系一般总是将知识领域划分成传统学科,再在其下进一步层层展开,建立以学科、专业为中心的分类体系。因此,在这种分类体系中,某一特定事物对象的有关信息资源往往是被分散的。为了适当解决这个问题,满足普通用户的使用需要,有些通用性的分类体系往往以事物对象为中心或事物对象与学科结合起来展开类目体系,这种方法在网络分类体系中有较为普遍的应用。

(3) 采用一定的标记符号作为排序工具。

标记符号是现代文献分类法必不可少的组成部分,用以表示类目的相对位置或相互关系。标记符号通常由一种或两种有序的标记系统中的符号(如数字或字母)组成,简短、明了,排序性好,将其用来作为信息资源的排检依据,可以充分发挥分类法进行资源组织和建立分类检索工具的作用。为了方便用户使用,网络分类检索工具在检索界面上一概并不显示其标记符号。

(4) 通过类目索引提供从字顺角度查找类目的途径。

为了克服因类目体系自身的特点而带来的检索困难,类目索引应运而生。类目索引提供了从语词字顺查找类目的便利方式。当然,通过这种方式查找到的不是一般的信息单元,而是同语词相对应的类目的分类号。也就是说,通过类目索引,可以确定特定主题在分类体系中的位置,进而方便、有效地使用类目体系。

信息资源分类的作用如下所述。

（1）进行资源组织。

即将分类法用于信息资源的组织。最常见的是用于信息资源的分类排架，分类排架是目前国际上大多数信息资源管理单位采用的方式。分类排架主要依据学科内容来组织信息资源，其显著的优点是：可以按照个体信息资源内容之间的关系将信息资源整体组织成一个有机系统。信息资源单位通常利用分类方法以开架的方式组织部分信息资源，为读者提供按照知识关系由此及彼地浏览、借阅的途径，对于读者来说，这种开架方式更加方便、直观，利用率很高。分类法不仅适用于传统文献，也适用于其他形式的信息资源组织，例如，可以直接用来组织数字资源，可以作为网络检索时分类收藏的依据，等等。

（2）建立分类检索工具。

即将分类法用于信息资源的揭示。分类检索工具主要包括分类目录、分类索引等，既可以是传统的卡片式、书本式等手工方式，也可以是现代的计算机检索系统。分类法是一种按照内容之间关系系统揭示信息资源的检索途径。进行资源组织和建立分类检索的工具，虽然都以分类法为依据，但是两者对分类法的要求有所区别，具体表现在三个方面：① 广度上，资源组织要求一种资源只能有一个排架号，而检索工具则要求多角度地揭示资源内容，因此一种资源可以有多个分类号；② 深度上，资源组织要求号码简短，便于查阅，分类相对要求粗略，而检索工具则要求类分详细，以加强专指度，提高检准率；③ 组织方式上，资源组织要求类目间采取单线排列方式，不能随意更换，而检索工具排列较灵活，类目间位置可以适当变动。

（3）分类统计。

分类统计是进行信息资源管理和利用的基本手段。信息资源管理部门可以通过分类统计，及时准确地掌握本单位各学科门类信息资源的配置和流通状况，发现各个知识门类信息资源存在的问题，进而深入了解用户对不同领域信息资源的需求，以便在逐类进行科学分析的基础之上有针对性地解决问题，有效地开展信息资源服务工作。

（4）兼容工具。

一般来说，比较有影响的大型综合性文献分类法往往会在若干个检索系统中同时使用，成为理想的跨库检索的工具；另外，作为系统揭示的工具，分类具有按照一定的知识结构组织信息资源的独特作用，利用分类语言可以对网上数据库的相关文献进行整合或以分类语言作为媒介，实现在不同检索工具之间进行转换的目的。

2. 分类法的类型

要从分类的角度准确、一致、有效地组织和揭示信息资源，规范、权威的工具是必不可少的，这个工具就是信息资源归类时需要遵循的类目体系，即信息资源分类法。

信息资源分类法，是按照类目之间关系组织起来的，并配有一定标记符号的分类信息资源的工具。信息资源分类法是聚类的结果，是根据分类需要预先建立的整体类目体系，是进行分类工作的依据和规范，是对数量巨大的信息资源进行分类组织的极其关键的因素。

从编制方式的角度来划分，信息资源分类法通常被分为三种类型，即等级列举式、分面组配式、列举—组配式。

1) 等级列举式分类法

等级列举式分类法是一种传统的分类法类型，是将所有的类目组织成一个等级系统，并且采用尽量列举的方式编制的分类法。这种分类法通常将类目体系组织成一个树形结构，按照划分的层次，逐级列出详尽的专指类目，并在以线性形式显示时，以缩格表示类目的等级关系。由于这种分类法通常是依据传统的知识分类体系编制的，所以人们习惯上也将其称为体

系分类法。从理论上说,这种分类体系可以不断递分下去至无穷,人们也称其为穷举式分类法或枚举式分类法。

等级列举式分类法的优点是:

- (1) 类目体系结构显示直观,符合大众思维习惯,容易掌握和使用;
- (2) 类目体系展开比较系统,并具有一定的伸缩性,使用过程中可以根据具体需要适当调整类目等级;

(3) 标记系统简单明了,既适用于进行资源组织,又适用于建立分类检索工具。

等级列举式分类法的不足是:

- (1) 类目展开方式相对单一,类目数量相对有限,类间组配能力较弱,无法完全满足确切分类的需要,无法充分揭示信息资源中大量存在的细小专深主题;
- (2) 类目体系结构相对固定,不便于根据需要随时改变、调整检索途径,不能进行多角度检索,也无法随时增加新类目以与科学发展保持同步;
- (3) 尽量列举类目的方式使得类表篇幅较大,增加了类表管理工作的难度。

等级列举式分类法是目前国内外使用得最普遍的分类法形式,比较著名的等级列举式分类法有:国外的《杜威十进分类法》(简称《杜威法》或 DDC, Dewey Decimal Classification)、《美国国会图书馆图书分类法》(简称《国会法》或 LCC, Library of Congress Classification)等;我国的《中国图书馆分类法》(简称《中图法》)等。

2) 分面组配式分类法

分面组配式分类法是依据概念的分析与综合原理,将概括信息资源内容与事物的主题概念组成“分面—亚面—类目”的结构体系,通过各分面内类目之间的组配来表达信息资源主题的一种信息资源分类法,也称为组配分类法、分析—综合分类法。分面组配式分类法的最典型的代表是印度著名图书馆学家阮冈纳赞(S. R. Ranganathan, 1892—1972)所创制的《冒号分类法》(Colon Classification)。

与等级列举式分类法相比,分面组配式分类法放弃了详细列举类目体系的做法,采用以简单概念组成复合类目的方式。

分面组配式分类法的核心是分面。分面又称组面,简称面,所谓面就是按某种分类标准(分类特征)产生出来的一组类目。

一个主题往往可以按照多个标准进行划分,形成多个分面。分面的依据主要是主题的范畴体系。进行分面分类,不仅要确定事物的主题,而且需要明确该主题所指向的范畴。在主题指向范畴的理解问题上,当前还存在着不尽相同的观点。最著名的、最有代表性的观点是印度图书馆学家阮冈纳赞的范畴理论,该理论主要包括如下5个方面:

- (1) 本体(Personality),是最主要的范畴,它是主题的整体,表示的是一种属性;
- (2) 物质(Matter),主要包括材料、载体等;
- (3) 动力或动因(Energy),该范畴涉及的方面比较多,如形态、功能、问题、方法、操作、处理及技术等;
- (4) 空间(Space),指事物存在或发生的地点;
- (5) 时间(Time),指事物存在或发生的时期。

在分面分析的过程中,多数情况下人们所遵循的就是上述这5个范畴。

由于在对信息资源进行分类的过程中,信息资源所涉及的主题往往不只是一个,在多主题的情况下,就包括多个组面。为了明确区分不同的组面,分面组配式分类法采用不同的组面区分符号。例如,《冒号分类法》中使用的分面符号如表3.2所示。

表 3.2 《冒号分类法》中使用的分面符号

基本范畴	组配符号	分面符号
本体 (Personality)	, (逗号)	(P)
物质 (Matter)	: (分号)	(M)
动力 (Energy)	: (冒号)	(E)
空间 (Space)	• (圆点)	(S)
时间 (Time)	' (反向逗号)	(T)

在明确了分面符号后, 还要注意分面组配式分类法的组配公式。组配公式是依据分面标记时所遵循的分面固定次序对各分面进行依次排列, 并赋予相应的分面组配符号而形成的。

由于知识门类的差异, 表现在相应的类目上的分面公式有所不同。因此, 分面组配式分类法往往都对各个类目规定了相应的分面组配公式。

在分面组配式分类法中, 还有一个重要概念, 即分面分析。分面分析实际上是主题分析的一种方法, 一般是指将一个主题分析为若干个分面。通常将编制组配分类法时的分面分析称为类目分面分析, 它是为建立分面结构的分类体系, 将特定领域的主题概念分析、归纳为一个一个分面的过程。

分面组配式分类法的优点是:

(1) 类表组配能力强, 标引结果专指性高, 可以通过基本概念的组配, 充分揭示信息资源中的复合主题;

(2) 标记表达性强, 便于根据不同需要调整组配次序, 实行轮排, 从不同角度提供检索途径;

(3) 较强的组配能力可以满足不断产生的新主题及复杂主题的标引需要, 与科学的发展保持同步;

(4) 与体系分类法相比, 类表的篇幅较小, 为分类法的管理、增补、修订等工作的开展提供了便利。

分面组配式分类法的不足是:

(1) 分面类表的类目体系是隐含的, 与体系分类法相比直观性不强;

(2) 分类检索工具中的类目是根据组配方式建立的, 导致各个学科门类的类目数量的分布往往缺乏均衡性;

(3) 组配方式及规则较复杂, 标引难度较大, 对分类标引人员的专业素养有较高的要求;

(4) 分面标记的构成相对比较复杂, 号码冗长, 在分类实践中, 不适用于组织信息资源排架, 主要应用于组织分类检索工具。

3) 列举—组配式分类法

列举—组配式分类法又称为半分面分类法, 是在等级列举式的详尽类表的基础上, 广泛采用各种组配方式的分类法。

列举—组配式分类法兼有前面两种分类法的特点。

列举—组配式分类法的优点是: 以列举式类表为基础, 具有一定的直观性, 同时广泛采用组配方法, 基本上可以达到与分面组配式类表同等的标引水平。

列举—组配式分类法的不足是: 在相关类目的修订或改进方面需要投入大量的资源和精力, 而且, 实现类目之间的组配时, 需要使用分面组配式分类法的多种辅助符号或号码进行标记, 最终导致分类标引方面的标记程序比较复杂, 标记符号也显得冗长。

3.2.2 分类法的编制

1. 分类法的结构体系

分类法的结构体系一般由类目体系、标记符号、说明与注释、类目索引四部分组成。

1) 类目体系

类目体系是根据类目内在关系和一定原则建立起来的类目集合,是分类法的主体,是分类语言进行词汇控制的主要依据。

类目体系一般是以知识分类为基础,按照信息资源分类的实际需要而建立的。类目体系由主表和复分表构成。

主表一般在基本部类的基础上,由基本大类、简表和详表等构成。

基本部类是为分类法的合理展开对知识范畴所做的最概括、最本质的划分,是信息资源分类法的纲目,所以,有的分类法将基本部类称为大纲。基本部类贯穿于主表之中。

基本部类的排列次序称为基本序列。对基本部类的划分及其序列的确定,是分类法展开的基础,通常应根据分类法的性质和使用需要,依据一定的知识分类体系确定。

我国文献分类法的基本部类,一般是依据对知识领域整体关系的了解,首先将知识门类划分为哲学、社会科学、自然科学三大部类,同时再根据马列主义、毛泽东思想的指导作用和文献分类本身的需要,增设马列主义、毛泽东思想和综合性图书两大部类,构成五大部类。

国外文献分类法的基本部类,通常也是在对学科领域及其关系了解的基础上确定的。例如《杜威十进分类法》的基本大类,是在理性知识、想象知识和记忆知识三大部类的基础上展开的;《冒号分类法》各主题领域的类表,是在自然科学、人文科学、社会科学的基础上设置的;等等。

由于基本部类的划分涉及分类体系的整体展开方式,所以受到分类界的广泛关注。但是,分类表中一般不单独列出基本部类,也并不是所有的分类法都明确确定了基本部类。

基本大类是分类法的基本大纲,是分类体系展开的起点,也就是通常说的第一级类目,它也是分类法整体框架的体现。

基本大类一般是在基本部类的基础上,根据知识门类发展情况和信息资源标引需要确定的。基本大类主要涉及类目设置和排列次序两个方面的问题。

(1) 基本大类的设置。

基本大类的设置,通常是与检索工具的性质、学科发展状况及信息资源数量等密切联系的。

传统的文献分类法一般均以学科为中心设置基本大类,构成从学科角度展开的大类体系。早期的文献分类法,由于受当时学科发展状况的局限,基本大类的设置比较概括,如《杜威十进分类法》,只设置了10个基本大类;其后编制的分类法逐步增加了大类数量,一般均将基本大类保持在20个左右,如《国会法》21个、《布立斯书目分类法》22个、《中图法》22个、《科图法》25个等。《冒号分类法》的基本大类曾经多达42个,不利于对类表的整体把握和标记分配,因此,第七版将基本大类减少为26个。而且,在同一部分类法中,各个基本大类的规模也比较均衡。

值得一提的是,与传统分类法不同,指南型网络分类工具的大类设置放弃了以学科为中心确定类目结构的传统,采用以主题为中心或主题与学科相结合的两种设类方式,实现了直接性与通用性的结合。基本大类的数量一般保持在14~20个之间。

(2) 基本大类的排列次序。

无论是国内还是国外的文献分类法,都比较重视对各个门类之间关系的揭示。

国外文献分类法一般根据从总到分的次序来序列类目，通常将综合性大类放在各个大类之首，其后再依次序列其他大类。

早期的《杜威十进分类法》，因为没有在大类序列中合理揭示诸如文学、语言学等大类之间的联系，受到了广泛的批评；在《杜威十进分类法》之后产生的《美国国会图书馆图书分类法》、《布立斯书目分类法》等，在大类排列上大多比较重视对大类之间联系的揭示，一般都将内容相关的门类集中设置，其中尤其以《布立斯书目分类法》的大类序列最为典型。

当然，国外不同分类法之间的基本大类也存在着明显的差异。例如，《杜威十进分类法》共有 10 个基本大类，其基本序列为：

- 综合性图书；
- 哲学与相关学科；
- 宗教；
- 社会科学；
- 语言；
- 纯科学；
- 技术（应用科学）；
- 艺术；
- 文学；
- 普通地理与历史及辅助。

由上述基本大类的序列可以发现，该分类法将综合性图书列在所有基本大类的首位，而将文学等类目列在纯科学之后。这个序列与我国的几部主要分类法的基本大类的序列存在很大的差异，究其原因在于：杜威吸收了英国哲学家培根的知识分类的原则与思想，并将其进行倒转，建立了“倒转培根法”的分类思想，这从基本大类中的哲学、艺术、文学等类目的位置顺序中可以清楚地看出来。

再如，《美国国会图书馆图书分类法》（LCC）的基本大类的序列与其他分类法相比，存在一些特殊的地方。在 LCC 的基本大类体系中，历史类目占有 3 个位置，军事科学类目之后紧跟的是海军科学，但在海军科学之后并没有顺理成章地列出陆军科学和空军科学类目，因此很难判断该分类法的基本大类中的分类思想和原则。但事实上，LCC 的分类体系并不是通常意义上的知识分类，因此，并不注重追求系统本身的严谨性，而是根据文献的收藏情况来建立分类法的类目体系。同样，在基本大类中的历史类，E-F 也主要是美洲史，这再次证明了文献情况是制约基本大类建立的主要条件。这个原则就是英国著名的文献分类学家赫尔姆（E.W.Hulme）所概括的“文献保证”原则。

我国文献分类法大类的序列，除了将马列主义、毛泽东思想设置为第一个大类等比较特殊的情况以外，在大类的排列上，一般都按照从总到分的原则，根据各个大类之间的学科关系来确定排列次序。例如，《中图法》在其大类序列中，将自然科学部门的大类按照从简单到复杂、低级到高级、理论到应用的次序排列；在社会科学门类中，将政治、法律、军事等上层建筑的类目，语言、文学、艺术等内容密切联系的类目，以接近的方式设类，以便使其可以较好地反映相关门类之间的联系：

- 马克思主义、列宁主义、毛泽东思想、邓小平理论；
- 哲学、宗教；
- 社会科学总论；
- 政治、法律；
- 军事；

经济;
文化、科学、教育、体育;
语言、文字;
文学;
艺术;
历史、地理;
自然科学总论;
数理科学和化学;
天文学、地球科学;
生物科学;
医药、卫生;
农业科学;
工业技术;
交通运输;
航空、航天;
环境科学、安全科学;
综合性图书。

与传统分类法不同,网络分类工具的大类多数未按照类目之间关系序列。英文网络分类体系大都以字顺方式排列基本大类,我国的网络分类工具则多数依据检索频率等因素来序列类目。这类方式方便类目的调整和增补,具有一定的实用性,但不符合分类法相关揭示的基本要求,逻辑性和系统性相对较弱。

简表也称为基本类目表或主要类目表,在分类法的使用中担负着承上启下的作用,是由分类表的基本大类的进一步区分而形成的简单类目体系,一般由二级类目或三级类目构成。

分面分类法通常在基本大类下列出基本类和分面公式,以帮助用户了解各基本类下涉及的类目范畴,其作用类似等级列举式分类法的简表。

简表主要有三个作用:①可以帮助用户迅速了解整个分类法的概况,引导进一步使用详表;②在不需要使用详表的情况下,可以用来进行概略分类,如中小型文献单位可以直接使用简表,专业文献单位可以用简表标引非专业信息资源;③便于用户进行较大范围信息资源的检索。

详表是由在简表的基础上进一步展开的不同级别的类目组成的类目表,是分类法的主体和正文,是类分信息资源的真正依据。

在等级列举式分类法中,详表由逐一列举的子目构成,并通过并列或缩格等方式显示其并列或等级关系。

在组配式分类法中,详表是由按范畴设置的基本概念表组成的,可以依据标引规则,以组配方式标引信息资源的主题内容。

复分表就是将主表中按相同标准划分某些类所产生的一系列相同子目抽取出来,配以特定号码,单独编制成表,供主表有关类目进一步复分用的类目表。

复分表也称附表、辅助表、副表、共性区分表,是分类法的重要组成部分。

类目表展开时,主表中有不少类目在进一步区分的过程中往往采用同一区分标准,区分出来的各个子目又大致相同。类目表展开时,主表中有不少类目在进一步区分的过程中往往需要采用相同的划分标准,而区分出的子目又大致相同。

例如,“文学”、“法律”、“体育”等大类和各专门学科总论性类目又可以进一步区分出

以下的子目：

理论与方法论；

现状及发展；

机构、团体、会议。

又如，“世界历史”与“文学史”，既可以按地域区分为下列类目。

世界历史：

中国历史；

朝鲜历史；

日本历史；

英国历史；

.....

文学史：

中国文学史；

朝鲜文学史；

日本文学史；

英国文学史；

.....

又可以按时代区分出下列类目。

世界历史：

古代史；

中世纪史；

近代史；

现代史；

.....

文学史：

古代文学史；

中世纪文学史；

近代文学史；

现代文学史；

.....

事实上，不仅“世界历史”和“文学史”需要按地域和时代进一步区分，许多文献都有可能或有必要按地域、时代、民族、语种、著作类型、体裁、学科、专业、事物的总论性问题等标准加以细分。

复分表主要有如下3个作用。

(1) 缩小类表的篇幅。

因为对于带有共性的问题，在任何一部分类法中往往涉及的类目很多，如果在这些类目的下面都设置相同的细分类目，势必造成主表的篇幅过于膨胀，因此，就这些共性问题制作成单独的表，采用同一个标准，会节省许多篇幅，通过使用附表，可以使类表在较小篇幅的情况下达到较大的细分程度。对于已经编制附表的共性类目，在需要进一步揭示时，不必在类表中重复设置子目，只要规定使用附表就可以了。

(2) 加强类表的灵活性。

灵活性包括两个方面：一方面，对于不同的单位，可以根据自身的具体情况来灵活掌握复分原则；另一方面，就分类法本身来说，灵活性主要体现在类目的伸缩性方面，可以根据实际使用的需要，通过在分类体系展开中增加或减少附表的使用，调整细分程度，增强类表的灵活性。

(3) 增强类表的规律性。

采用统一方式编列共性子目、配置号码，有助于使类目体系的列举更加一致，使分类号的结构具有一定的规律性，有利于揭示文献中复合主题的各主题因素，增加类目的助记性。

按其使用范围，复分表可分为通用复分表和专类复分表两种。此外，对部分共性子目采用的仿分形式也具有与复分表相似的作用。

通用复分表又称共同区分表，是一种供主表各大类有关类目共同使用的表，通常在类表的前部或后部集中编列。

综合性的文献分类法中常用的通用复分表有：标准复分表、地区复分表、时代复分表等。

标准复分表在国内亦称总论复分表，是一种通用主题和信息资源类型的复分表，通常收

入适用于整个分类体系的通用性主题,如原理、方法、历史、现状、机关、团体、会议等,以及常用文献形式,包括媒体类型、文献类型等各种有关类目。

地区复分表收录自然地理、各国政区等类目。

时代复分表收入时间、历史、时代等的有关类目。

此外,一些分类法还设有语种复分表、语言形式复分表、世界种族与民族复分表、人物表、材料复分表、环境复分表等多种类型。

以《中图法》为例,在主表之后,共设置了8个通用复分表:

总论复分表;

世界地区表;

中国地区表;

国际时代表;

中国时代表;

世界种族与民族表;

中国民族表;

通用时间、地点表。

专类复分表是一种只限于在某一基本大类或专门领域使用的复分表,一般设置于相应类目之下。

等级列举式分类法为缩小类表的篇幅,往往根据各类的需要设置多种专类复分表。相对于通用复分表而言,专类复分表的使用范围比较狭窄,它只供某大类中的有关类目复分时使用,其他类不能依此表复分。如《杜威十进分类法》的文学复分表是供文学大类中有关类目复分用的。又如《中图法》(第四版)共设有58个专类复分表。分面分类法由于分面序列本身具有复分表的同等作用,因此通常不再设置专类复分表。

以《中图法》为例,在“J62/628 各种器乐理论及演奏法”下面就设置了一个可供从“J62”到“J628”这一区间的各级类目使用的专类复分表。

仿分也是一种以统一的方式处理共性子目的方法。

在类表中,一组性质相近的类目出现相同的子目时,为了简化类表的编制,增加类目设置的规律性,通常在总论性类目或在前类目下详列子目,指明供有关类目仿照复分。这种利用某一类的子目作为进一步区分依据的类目处理方法即称为仿分。

仿分有类似专类复分表的作用。各种分类法,包括分面组配式分类法也常采用这一形式。有些分类法,如《杜威十进分类法》,还规定某些类目可以仿基本大类等进行区分。

以《中国图书馆分类法》为例,在“F593/597 各国旅游事业”下面注释:依世界地区表分,再仿F592分。意思就是,从“F593”到“F597”区间所包括的所有国家的旅游事业的类目,都可以仿照“F592 中国旅游事业”下面已经详细展开的类目体系来进一步展开。

复分表是一种辅助区分的手段,使用时一般应注意:① 除有明确规定外,一般不得单独使用,必须结合主表类目使用;② 是否使用复分表,应按表中有关的规定进行,通常可根据复分表中的说明及类目下的注释确定;③ 各文献单位可以根据需要对复分表的使用加以调整或限定,但一旦确定,就应严格遵守,不得随意变动,以保持复分表使用的一致性;④ 由于复分表使用意味着新的分类成分的插入,一般还应注意标记配置方面的有关规定。

2) 标记符号

标记符号将在本章的有关内容中介绍,在此不赘述。

3) 说明与注释

说明与注释是分类法中帮助用户了解和使用分类体系的重要组成部分,一般用来说明类

表的编制原则、类目体系的特点及使用方法等。说明与注释通常包括编制说明、大类说明和类目注释等三种形式。

编制说明也称为序论，是对分类法分类体系整体情况的纲要性说明。编制说明的内容一般应包括该分类法所依据的编制原则、分类体系的特点、号码制度及基本使用方法等。

例如，《中图法》第一版的“编制说明”的主要内容包括：编制原则、体系结构、关于分类体系的几点说明、标记符号。第四版的“编制说明”的主要内容包括：指导思想、修订原则、修订工作进程、修订重点及特色。

通过阅读编制说明，可以从整体上了解分类法类目体系的特点。

大类说明是对分类法各个基本大类结构特点和标引规则的纲要性说明。大类说明一般由该分类法中特定基本大类的内容范围、类目体系或分面结构的特点及该基本大类的主要标引规则等内容组成。通过阅读大类说明，可以了解各基本大类的结构、范围、与相关知识门类的关系及分类标引的要点等，所以，大类说明是掌握各基本大类分类方法的重要依据。

类目注释是对分类法中类目名称的补充说明。类目注释主要用来对具体类目的含义、范围、与其他类目的关系及具体的使用方法等做进一步的说明，有利于对类目的使用进行规范性控制，保证分类标引具有准确性和一致性。

下面以《中图法》为例，简单概括其类目注释的主要类型。

第一，规定了类目的含义和内容范围。

例如：Q813.2 细胞融合工程

用自然或人工的方法，使两个或几个不同的细胞融合成一个细胞的过程。

学习杂交瘤技术、原生质体融合、单克隆抗体等入此。

再如：I210.96 鲁迅思想的学习和研究

学习鲁迅精神的著作入此。

第二，揭示了某类与相关类的关系，如指明交替类目、类目参照及参考有关类的标引方法等。

例如：[O341] 材料力学
宜入 TB301。

又如：TH744.5 激光仪器
参见 TN248。

再如：A849.18 语录
见 A18 注。

第三，指明了具体的分类标引方法如类目复分、仿分，说明了特殊分类规则和配号方法等。

例如：C97 劳动科学
依总论复分表分，-0 理论与方法论所属类目入 C970。

又如：F593/597 各国旅游事业
依世界地区表分，再仿 F592 分。

再如：U448 各种桥梁
如果遇多主题因素文献，入前面编列的类。例如《铁路钢筋混凝土斜拉桥》入 U448.13。

第四，规定了同类文献的排列方法。

例如：G812.17 地方体育运动组织

体育俱乐部入此。

依中国地区表分,按名称字顺排。

再如: P416.3 高空观测记录

依世界地区表分,再按年代排。

第五,说明了具体类目的修订情况。

例如: {P413.1} 计算单位

<停用; 4 版改入 P413>

再如: J211.22 经济、文化

科学、教育、体育、卫生等入此。

<3 版类名: 生产建设>

此外,《中图法》(第四版)还以突出的方式,通过注释对一些特殊分类规定加以说明。

作为分类法必不可少的重要组成部分,说明与注释对于增强分类法的易用性和标引的一致性、提高分类标引的质量具有关键意义,所以,随着现代分类法的发展和各种分类技术的应用,它在国内外分类法中的运用越来越得到重视,一些著名的文献分类法如《中图法》、《杜威法》等都加强了对说明与注释的编制。

由于分类体系涉及的知识门类多、范围广,类目之间关系错综复杂,对于一些细小专深主题的查找往往比较困难,既需要具有相关门类的专业知识,又必须对分类体系展开的规律和类目设置的特点有比较深入的了解。即使如此,对于一些同时涉及若干个门类的复合主题,还是需要几经周折才能找到最确切的类目,进而达到正确标引的目的。

4) 类目索引

而类目索引,也称分类表索引,是从类目名称字顺途径查找相应分类号的工具,为分类法的有效使用提供了便利。

类目索引主要有两个作用:① 通过将分类体系的系统排列转变成字顺排列,可以从表达主题的词语出发找到相应的分类号,克服了类目查找的困难;② 便于用户查找分类表中被分散在各个学科门类的有关同一事物的类目及分类表中未列出的有关新概念。

类目索引主要包括直接索引、相关索引和主题词索引三种类型。

直接索引是一种直接通过类名或同义词查找对应类目的索引。

直接索引通常将分类表中所有类目及其注释中的有关主题概念,按照名称的字顺排列,每个类目一般只按照分类表中的措词在索引中出现一次,在每条索引款目后注明相应的分类号,达到从类目名称查找对应分类号的目的。

直接索引的编制比较简单,功能也相对单一,索引款目缺少字面成族的机会,只能从一主题的语词出发查找对应类号,不能揭示与一个主题对象相关的类目之间的联系,难以反映复杂的专指主题,其效能不如相关索引。

相关索引是一种不仅可以从主题名称出发查找对应的类目,而且可以将被分类体系分散的该主题各方面的类目加以集中的工具。

相关索引除了把分类表中的全部类目及其注释中具有检索意义的主题概念按其字顺排列外,还采用一些综合材料的方法,将一个主题各个方面及被这一主题所规定的词,都集中在一个主题标目之下。

相关索引为美国图书馆学家杜威首创,最先用于《杜威法》索引的编制,因可以揭示分类法中的相关材料而得名,受到文献分类界的重视。为了将一个主题各个方面的类目集中在一起,这种索引除采用直接索引的方法外,还采用倒置、按学科集中等形式。

相关索引的编制比直接索引复杂,特别是在对其中字面不同的相关类目进行集中时,

有较大难度。相关索引的突出优点是：可以集中与一主题对象有关但在类目体系中被分散了的类目，有助于在标引和检索中扩大查找范围，对相关类进行揭示，其性能远远超过了直接索引。

主题词索引是一种将分类法与主题法结合的索引类型。

主题词索引是在分类主题一体化的背景下产生的。20 世纪 70 年代以后，分类主题一体化类表编制并广泛使用，一些分类法与主题法的有效结合，发展了标题索引、叙词索引等索引形式，提供了从主题词出发查找和使用类目的途径。

随着计算机在分类法编制中的应用，编制有详尽索引的机读分类表将成为主要分类工具。机读分类表的类目索引一般可以利用文本检索的各种方式对类目进行匹配查找，揭示充分，形式灵活。此外，这类索引收入的数据通常也远比书本式索引充分。

当然，类目索引虽然是分类法的重要组成部分，但只是辅助工具，不能直接以此作为分类的依据。在需要使用类目索引时，必须是在查核类表、明确某类在类目体系中的位置及其与相关类目的关系、了解该类的确切含义的基础上，才能最终归类。

2. 分类法类目体系的建立

类目体系建立的过程就是类目分析的过程。

所谓类目分析，是指在编制分类法时，对于要建立的类目体系中的类目进行严格分析的过程。

建立类目体系的方法有两种：① 归纳方法，即从个别到一般，根据个体属性的相同点集合成类，并依照这一方式，逐步将小类聚合成大类，建立起类目体系；② 划分方法，即从总到分，将若干个概略性类目作为一级类目，按照特定的分类标准逐级划分、层层展开，建立类目体系。

一般来说，在编制分类法的过程中，上述两种方法是结合使用的，往往在按照层层划分的方法展开类目体系的同时，有不少类目也采用了类聚的方法进行设置。考虑到国内多数分类法的类目体系属于逐级划分的等级列举式类型，而且长期以来拥有极其广泛的用户基础，所以，下面主要介绍采用从总到分的划分方法来建立类目体系的问题。

（1）首先是类目划分问题。

所谓类目划分，是指依据一定的属性或特征对类目的外延进行区分，生成一组子目的过程。

在对事物进行划分的过程中，用来作为划分依据的事物属性称为分类标准。分类标准有主要标准和辅助标准两种。

一般来说，以信息资源的本质属性即信息资源所体现的学科知识内容作为分类的主要标准，以信息资源的其他属性特征作为分类的辅助标准。

分类的实践证明，以本质属性即学科知识内容作为主要分类标准，具有科学认识上的意义与检索利用价值；而以其他显著属性特征作为辅助分类标准，也能满足用户的某些检索、利用要求。

（2）然后是引用次序问题。

引用次序，也称为组配次序，指复合主题标引和检索时，各个主题因素的组合次序。

在分类法中，引用次序是指类目划分标准使用的次序或不同分面的概念在组配时被引用的先后次序。在先组式检索工具中，引用次序具有决定复合主题类聚方式和排检位置的作用，是进行句法控制的重要手段。

采用明确的引用次序具有以下作用：

① 以用户使用需要作为组织检索系统的依据，提高检索系统的通用性；

② 保证复合主题处理的一致性,不因组配次序的不一致而影响相同主题信息资源的集中程度;

③ 保证分类标引结果含义明确,容易理解,避免出现错误的、无价值的组配结果,增强分类标引的准确性;

④ 使分类标引具有一定的规律性,提高分类检索系统的一致性和可预见性。

在不同的情况下,引用次序往往会表现为不同的形式。

在列举式分类法中,引用次序是指分类体系展开过程中划分标准使用的先后顺序;在分面组配式分类法中,引用次序则表现为复合主题中不同组面被引用的先后次序。

例如,在列举式分类法中,将“中国小说”这一主题按照“中国文学—小说作品”的次序组配,就是指分类体系把有关各国文学作品的信息资源按照“各国文学—作品”的次序加以引用。按照这一引用次序,分类体系首先按文学的国别对信息资源进行处理,将各种有关中国文学的信息资源集中在一起,不同国别的小说信息资源则被分散在各国文学之下。相反,如果按“小说作品—中国文学”的次序加以组织,就可以将不同国别的小说作品信息资源集中在一起,但关于中国各种体裁的文学作品的信息资源则被分散在各种不同体裁的文学作品之下。

在分类体系建立的过程中,选择什么属性作为分类标准、采用什么引用次序,都决定着类目体系展开的方式,会对分类体系的性能产生直接的影响。引用次序决定着类目体系以何种方式集中信息资源及提供何种检索途径,与分类体系的适用性密切相关。

在多数情况下,建立类目体系时,先采用主要标准,后利用辅助标准,但这也不是绝对的,主要标准并不严格地意味着必须要首先采用。关键的问题是,要根据用户的检索需求、结合主题领域的特点,确定适用的引用次序。为了同时兼顾各种不同引用次序的特点,满足从不同角度检索的需要,各种网络分类体系往往同时采用几种引用次序来揭示各类信息资源。

3. 分类法的标记系统

1) 标记符号

标记符号,亦称分类号,是分类法中用于标识类目的代号。

标记符号是现代分类法的重要组成部分,在一部分类法中,特定的标记符号与具体的类目之间存在着对应的关系。现代文献分类法正是以标记符号为中介,将分类体系有效地用于组织文献收藏和建立检索工具的。

分类体系是按照类目之间的关系确定的,而分类表中的类目数量众多、关系复杂,人们无法凭记忆来确定特定类目在整个分类体系中的相对位置。因此,要使一部分类法成为一个实用工具,必须用相应的符号系统来表示分类体系中众多类目的次序和关系。分类法的标记符号就是为了满足这一需要而设置的。

信息资源分类法的标记符号的作用有两个:① 固定类目次序;② 显示类目之间的关系。这是标记符号的积极作用。

当然,标记符号也会给分类体系带来限制:① 分类体系中的具体类目一旦配置了特定的标记符号,整个分类体系就要受到标记符号的约束,任何标记符号的变动都必须考虑到与其他已配置标记符号的关系;② 如果标记符号的配置方法不妥当,分类体系的扩充或调整就有可能受到标记符号系统的限制,从而影响分类体系的发展。

因此,标记符号的配置应当根据分类体系的特点和需要进行,并应力求将标记符号对分类体系的限制降到最低的限度。

按照号码组成成分,标记符号通常可分为单纯号码和混合号码两种。

单纯号码是指由一种具有固定次序的符号系统构成的号码。常用的有数字和字母两种。单纯数字号码顺序性强，易于排检，具有国际通用性，但基数较小，在采用表达性标记的情况下，当下位类的数量超过 10 个时，通常必须使用两位数字表示一次划分，从而会使号码变长。

字母标记基数大，在使用层累标记制的情况下，大多可以用一位字母表示一次划分，号码相对较为简短。字母标记制的不足是：其排检不如数字迅捷，使用也不如数字普遍。

目前流行的分类法中，使用单纯数字号码的较多，如我国的《科图法》、美国的《杜威法》等均为单纯数字标记；采用单纯字母标记的多为专业分类法，如英国的《伦敦教育分类法》等。

混合号码是指由两种或两种以上具有固定次序的符号系统构成的号码。通常由数字、字母结合使用，目的是取长补短，同时汲取两种符号的长处。

混合号码一般以字母标记基本大类或二级类，用以减少标记的位数，其余以数字为标记，必要时可再在标记末端适当采用字母标识。

我国的《中图法》、美国的《国会法》等大型文献分类法均采用混合号码。

单纯号码和混合号码一般还根据需要使用各种任意符号，如：“：”、“=”、“（）”等作为辅助符号，以增加标记的表达性。

概括地说，信息资源分类法的标记符号具有如下 4 种性能。

（1）容纳性。

容纳性又称扩充性，是指标记系统应具有根据类目体系的增补变动及时配予相应号码的能力。也就是说，标记符号不束缚类目体系的发展，能适应任何类系和类列的必要细分和增补，能为新出现的学科或主题配上符合类目体系要求的恰当的标记符号。

标记符号的容纳性，是由类目体系的动态性所决定的。

因为信息资源分类法往往与信息管理服务单位的服务工作密切相关，所以类目体系应保持相对的稳定性，但这种相对的稳定性是同科学知识分类体系相比较而言的。事实上，分类法的类目体系并不是一个绝对的、静止的封闭系统。

究其原因，一是科学技术的迅速发展，反映新学科、新事物、新问题的信息资源大量增长；二是人们对信息资源组织包括信息资源分类规律的认识逐渐深化。这两个因素使得信息资源分类法不断需要补充新类目，甚至局部地改变原有的类目体系，处于不断完善的动态之中。

而标记符号与类目体系具有密不可分的对应关系，因此，要使分类体系能满足信息组织的要求，必须不断对原有类目体系进行各种修订、增补和调整，使标记系统能根据类目体系发展的需要，在相应位置上以适用的方式对任何形式的增补、插入和细分加以标示。

增强标记符号容纳性的主要方法有：使用小数制，采用非等级性标记，在应用等级标记的情况下结合使用各种灵活标记形式及增设辅助符号等。

（2）表达性。

表达性主要包括两层含义：一是标记符号能够体现类目体系的结构特点，二是标记符号能够显示信息资源中各种学科主题之间或某学科主题中各个主题因素之间的相互关系。

信息资源分类法标记符号的表达性主要是通过采用分面标记制、层累标记制等来实现的。具体来说，目前分类法中使用的表达性标记主要有两种：一种仅能显示类目划分的等级，反映类目的从属或并列关系；另一种则能进一步显示复合主题的组配结构，表达主题间的各种联系。前者常在等级列举式分类法中使用，后者一般结合分面组配式分类法予以配置。

实现标记符号表达性的方法包括：以号码的位数表示类目划分的等级，以一定的辅助符

号或配号方式揭示类目所属的组面等。

在机检系统中,具有表达性的标记符号有助于对类目体系进行等级显示及组配检索等,因此,供机检使用的类表多采用等级标记。但是,标记符号的表达性也会带来一些负面的作用。例如,会增加标记符号的长度,使标记符号的成分复杂,影响标记符号的简明性和容纳性。

著名的《冒号分类法》采用了分面标记制,其标记符号不仅具有较强的容纳性,而且还具有显著的表达性。例如,用《冒号分类法》对主题“1950年印度经济状况”进行标引的结果是:X.44‘N5。其中,“X”表示经济,“44”表示印度,“‘N5”表示1950年。显而易见,从分类号的分段形式上,可以识别出构成该复杂主题概念的各个主题因素。

同样,著名的《国际十进分类法》属于体系—组配分类法,它的分类号也具有较强的表达性。其中,有些复分号既能连接于主类号之后,又能置于主类号之前或插入主类号中的适当位置,使信息管理单位在进行信息组织时,能够根据本单位的性质和用户的需要,局部地改变原有的分类体系,集中某些信息资源。例如,用《国际十进分类法》标引出的三个结果:

629.13(41) 航空工程中的英国概况

(41) 629.13 英国航空工程

629.1(41)3 运输工程中的英国航空工程

分解后可知,“629.1”表示“运输工程”,“3”表示“航空工程”,“(41)”表示“英国”。层累标记制的表达性更是显而易见的。

(3) 简明性。

简明性是指标记符号要简短明了、顺序性强、易读易记、便于排检,能够满足实际使用的需要。作为文献排架、分类检索依据的标记符号,必须避免号码冗长、成分复杂,否则会影响使用的效率。

在设计与配置标记符号时,应在保证标记符号具有较强的容纳性和一定的表达性的前提下,充分注意标记符号的简明性。常用的方法有:只使用一种有固定次序的符号系统,不用或少用辅助符号;选择基数比较大的符号系统,如字母的基数就比数字大;均衡地配置号码;适当降低类目揭示深度;采用非表达性标记或局部放弃表达性的标记等,如后面将具体介绍的“几种特殊的标记方法”。

(4) 助记性。

助记性通常是指对类表中同一含义的主题概念使用相同的号码进行配置。助记性要求配备的标记符号具有规律性及可理解性,有利于提高分类检索的速度与效果,也有利于信息资源实体的排架。助记性往往是与复分表、分面组配等形式相联系的。例如《国际十进分类法》中,“(4)”总是表示欧洲,“(51)”总是表示中国。《中图法》通常也在号码配置时对一些重复出现的概念以统一的方式配号,如以“0”表示理论,以“2”表示中国,以“9”表示历史等。通过号码配置中的规律性,提高标记符号的助记性。

一般来说,增强标记符号助记性的措施主要有如下几种:

第一,基本大类采用字母制,某些二级类目采用双字母制。例如《中图法》采用汉语拼音字母标记22个基本大类,用双位字母标记工业技术所属的16个二级类目,其助记性比采用单纯阿拉伯数字标记基本大类与某些特定二级类目要强得多;

第二,将分类表各类中的有关地域、民族、时代及学科、专业与事物的总论性问题等具有共性的子目集中起来,单独编列复分表,配以固定的号码,使同性类目的细分规范化,使分类号的结构具有一定的规律性;

第三,对具有相同区分标准的类目,在采用对应列类法的同时也采用对应编号法。

应该指出,上面介绍的标记符号4个方面的性能并不是孤立存在的,它们彼此之间具有相互联系、相互制约的关系。如果在编制和使用标记符号时,单纯、片面地追求某一方面的性能,就有可能降低其他方面的性能。

比较科学的做法是:依据信息资源分类法的主要功能,适当地对标记符号4个方面的性能加以协调。一般情况下,应在保证标记符号具有较强的容纳性的前提下,增强表达性、助记性与简短性,并尽量使四者协调起来。

2) 标记制度

标记制度是指分类号码的编制方法,即为类目编配号码的方式。

按照号码的组成方式,分类标记一般可以分为层累标记制、顺序标记制、顺序一层累标记制、分面标记制等基本类型。

层累标记制亦称等级标记制,是一种能够显示类目之间的等级关系和结构的标记制度。类目结构与类目关系的显示主要是通过号码与类目的对应性来体现的。一般是一级类目用一位号码、二级类目用两位号码,同位类再顺序配予号码。这种标记体系与分类法层层划分的类目体系相对应,因此,该标记制度不仅可以反映类目次序,而且能够体现分类法中的层次结构,可以根据标记的位数判断出类目的等级。网络分类法中为了便于扩展号码,往往也使用两位甚至三位号码表示一次划分。

层累标记制的优点是:可以揭示类目的等级结构,使用它,能够在机检系统中通过标记逐级显示分类体系。其不足是:如果类目划分等级较深,会造成号码过长;其次,如果同位类数量较多,超过号码的基数,就无法严格按等级编号;此外,在顺序配号的两个同位类之间出现新类时,号码的扩充也会出现问题。我国的《中图法》基本采用层累标记制,下面就是《中图法》中层累标记的例子:

G	文化、科学、教育、体育	(一级类目)
G8	体育	(二级类目)
G86	水上、冰上与雪上运动	(三级类目)
G861	水上运动	(四级类目)
G861.1	游泳	(五级类目)
G861.11	自由泳	(六级类目)
.....	

顺序标记制是一种只反映类目的先后顺序而不反映类目的层次结构的标记制度。这种标记制度中,仅从类号无法判断类目体系中类目之间的内在关系,该标记制度能够依据类目的数量较为平衡地进行号码分配,标记结果简短,容纳性强,适合文献排架,但最大的不足就是不能够揭示类目体系的结构,无法在机检系统中通过标记按等级显示机读文档。

顺序标记制通常分为字母顺序标记制和数字顺序标记制。在数字顺序标记制中一般又可以分为整数顺序标记制和小数顺序标记制。另外,还有字母与数字混合顺序标记制度,《美国国会图书馆图书分类法》就属于这种类型的标记制度。由于《美国国会图书馆图书分类法》的编制主要遵循的是“文献保证”原则,不重视类目体系中各类目之间的逻辑性,因此,在标记制度方面,一般只能采用顺序标记制度。例如:

L	教育
LB	教育理论和实践
LB2300—2330	总论
LB2331	专论的总类
LB2332	学术自由

LB2333	效率、评价
	工资、津贴
LB2334	综合著作和美国的著作
LB2335	其他国家 A-Z
LB2335.3	供给和要求. 周转率
LB2335.35	工作负担
LB2335.4	教育辅助
.....

顺序—层累标记制是一种顺序制与层累制相结合的标记制度。采用这种标记制度的目的是试图汲取顺序制与层累制的优点，使分类标记在性能方面既具有较强的简明性和容纳性，又能够保持一定的表达性。

我国的《科图法》的标记系统采用的就是顺序—层累标记制，如：

71	工程技术
75	金属学、物理冶金
75.6	铁碳合金
75.63	合金钢
75.633	工具合金钢
76	冶金学
78	机械工程、机械制造
79	各部门专用机械
又如：	
65	农业科学
66	农作物
66.1	谷类作物（粮食作物）
66.11	小麦
66.13	大麦、裸麦（元麦）
.....
69	畜牧、兽医、蚕蜂、水产
.....	
71	工程技术
72	能源学、动力工程
72.1	电能学
72.11	电的产生
72.111	直流电
72.112	交流电
.....

可以看出，《科图法》分类标记的前两位数字以顺序方式配号，用于标示基本大类、二级类甚至三级类；两位数字之后，用小圆点隔开，其后基本使用层累标记制。

分面标记制是能够反映类目标组配结构的标记制度。该标记制度通常采用一定的标记符号和特有的组配方式来表示一个复杂主题的各要素所属的分面。该标记制度既能够反映类目的等级和先后顺序，更能揭示类目的分面结构。常见的分面标记采用分段组合方式，即通过分面符号的使用，使标记成为一种由若干具有独立性的节段构成的组合号码，可用来进行轮排

或以组配方式检索。

例如,用《冒号分类法》对主题“日本 70 年代水稻疾病防治”进行标引,结果为:

J, 381; 421: 5. 42 ‘N7

其中,“J”表示基本类农业,“381”表示水稻,“421”表示疾病,“5”表示防治,“. 42”表示日本,“‘N7”表示 20 世纪 70 年代,辅助符号“,”、“;”、“:”、“.”、“‘”为分面指示符,分别表示本体、材料、动力、空间、时间五个不同分面。

又如,同样用《冒号分类法》对主题“20 世纪 90 年代中国儿童图书馆图书分类工作”进行标引,结果为:

2, 61; 43: 51. 41 ‘N9

其中,“2”表示基本类图书馆学,“61”表示儿童图书馆,“43”表示图书,“51”表示分类,“. 41”表示中国,“‘N9”表示 20 世纪 90 年代,辅助符号“,”、“;”、“:”、“.”、“‘”为分面指示符,分别表示本体、材料、动力、空间、时间五个不同分面。

分面标记制的优点是,具有较强的表达性和可组配性,不仅能够充分揭示信息资源主题,还可以进行轮排和组配检索。其不足是号码成分比较复杂,标记结果冗长,排序性能差,不适合用来组织信息资源排架。

3) 几种特殊的标记方法

为了使标记系统具有较好的性能,在具有容纳性的同时保持一定的表达性和简明性,分类标记一般还采用以下各种标记技术,以更好地适应类目体系及其发展的需要。

(1) 八分法。

八分法又称“扩九法”。八分法是在采用层累数字标记的情况下,当同位类超过 10 个又不足 18 个时,前 9 位以 0~8 表示,8 后面的标记用两位数字表示一次划分,这种标记技术用于解决同位类的号码配置问题。下面是《中图法》中运用八分法的例子。

Q429	末梢神经系统的生理
Q429.1	嗅神经
Q429.2	视神经
Q429.3	动眼神经
Q429.4	滑车神经
Q429.5	外展神经
Q429.6	三叉神经
Q429.7	面神经
Q429.8	听神经
Q429.91	舌咽神经
Q429.92	迷走神经
Q429.93	副神经
Q429.94	舌下神经

“Q429 末梢神经系统的生理”下展开的一组同位类中,从“Q429.8 听神经”到“Q429.94 舌下神经”的类号就采用了八分法。

八分法虽然在数字的位数上不能体现所有同位类之间的并列关系,但由于它用“9”作为空位号来扩展同位类的号码,因此仍能有规律地表示各个同位类。

八分法从理论上说可以不断地用空位号“9”无限地扩展号码,但会使号码越来越长,影响了标记符号的简短性、表达性及助记性三个性能的发挥。因此,在使用八分法时,一般最多只采用两次“八分”,如果同位类数量过多,通常适宜采用双位制标记技术。

(2) 双位制。

双位制又称百分法。双位制是在采用层累数字标记的情况下,当同位类超过18个时,直接以两位数字表示一次划分,以解决号码的扩充问题。双位制是为了增强类列的容纳性而采用的一种标记技术。下面是《中图法》中的例子,虽然不是严格的双位制,但基本上属于这一类型。

I236	地方剧
I236.21	天津市地方剧
I236.22	河北省地方剧
I236.25	山西省地方剧
I236.26	内蒙古自治区地方剧
I236.3	东北地区地方剧
I236.41	陕西省地方剧
I236.42	甘肃省地方剧
I236.43	宁夏回族自治区地方剧
I236.44	青海省地方剧
I236.45	新疆维吾尔自治区地方剧
I236.51	上海市地方剧
I236.52	山东省地方剧
I236.53	江苏省地方剧
I236.54	安徽省地方剧
I236.55	浙江省地方剧
I236.56	江西省地方剧
I236.57	福建省地方剧
I236.58	台湾省地方剧
I236.61	河南省地方剧
I236.63	湖北省地方剧
I236.64	湖南省地方剧
I236.65	广东省地方剧
I236.658	香港地方剧
I236.659	澳门地方剧
I236.66	海南省地方剧
I236.67	广西壮族自治区地方剧
I236.71	四川省地方剧
I236.719	重庆市地方剧
I236.73	贵州省地方剧
I236.74	云南省地方剧
I236.75	西藏自治区地方剧

双位制从理论上说可以编100个号码,但通常不使用带“0”的双位数字,因此从号码11~99,可以容纳81个同位类。

双位制的容纳性很强,但简明性、表达性及助记性都较弱。

(3) 借号法。

借号法是在采用层累标记制的情况下,为了增强类列的容纳性而采用的借用上位类或下

位类类号的一种标记技术。通常在上一级号码较宽裕时，用上级号码标示下级类目，或在下级类目超过10个但超过不多时，根据情况，借用“9”以外的下级号码进行扩充，使号码配置更加灵活。下面是《中图法》中借上位类类号的例子。

K235	三国、晋、南北朝（220—589）
K236	三国（220—280）
.....	
K237	晋（265—420）
.....	
K238	十六国（304—439）
K239	南北朝（420—589）
.....	

其中，“K236 三国”、“K237 晋”、“K238 十六国”、“K239 南北朝”四个类的类号借用了其上位类“K235 三国、晋、南北朝”的类号。

（4）预留空号法。

预留空号法是指根据学科发展情况和类目设置的可能，在配号时预先留下一些空号，供类目增补时使用的一种标记技术。空号法在分类法中运用得很普遍。下面是《中图法》中预留空号法的例子。

E07	军事管理学
E071	军制学
E072	军队指挥学
E073	军事教育学
E075	后勤学

（5）双位加点法。

双位加点法是一种通过在两位号码后加点表示一次划分的扩号方法。我国的《中国人民大学图书馆图书分类法》采用了这种标记技术。下面是《人大法》（第6版）的基本大类，其中就采用了双位加点法。

1	马克思主义、列宁主义、毛泽东思想
2	哲学
3	社会科学、政治
4	经济
5	军事
6	法律
7	文化、教育、科学、体育
8	艺术
9	语言、文字
10.	文学
11.	历史
12.	地理
13.	自然科学
14.	医药、卫生
15.	工程技术
16.	农业科学技术
17.	综合性科学、综合性图书

其中,从“10.”到“17.”为双位加点法,其代表的类目级次与“1”到“9”相同,均为一级类目。

(6) 字母标记法。

字母标记法是在使用数字标记的情况下,直接以类名的首字母为标记,标示下一级类目的一种标记技术。这种方式一般只用来标示类表的最后一级类目。美国《国会法》常采用这一方法。我国的《中图法》也部分使用了这种方法。下面是《中图法》中字母标记法的例子。

TP312 程序语言、算法语言

依语言名称的前两位英文字母区分,并按字母序列排,若程序语言名称的前两位字母相同时,则取第三位字母,依此类推。例如,ALGOL 语言为 TP312AL,Java 语言为 TP312JA,TP312 排在 TP312JA 之前。

(7) 对应编号法。

对应编号法是按照类目设置的规律统一配置对应号码,使标记具有规律性、一致性,方便用户使用的一种标记技术。换句话说,对应编号法是针对具有相同区分标准的类目,使其相应部分的号码趋于一致,以增强标记符号易记性的一种编号方法。下面是《中图法》中对对应编号法的例子。

H1	汉语
H11	语音
H12	文字学
H13	语义、词汇、词义(训诂学)
H14	语法
H15	写作、修辞
H159	翻译
H16	字书、字典、词典
H17	方言
H19	汉语教学
H31	英语
H311	语音
H312	文字
H313	语义、词汇、词义
H314	语法
H315	写作、修辞
H315.9	翻译
H316	词典
H317	方言
H319	语文教学

其中,“H1 汉语”类下从“H11 语音”到“H19 汉语教学”与“H31 英语”类下从“H311 语音”到“H319 语文教学”两组类号就存在对应关系。

显然,对应编号法突出了标记符号的表达性和助记性两种性能。

3.2.3 国内外常用分类法介绍

下面概要介绍国外四部著名的分类法《杜威十进分类法》(DDC)、《国际十进分类法》(UDC)、《美国国会图书馆图书分类法》(LCC)和《冒号分类法》(CC),以及我国最著名的

分类法《中国图书馆分类法》。其中，DDC、UDC 和 LCC 是目前在国外使用得最广泛的文献分类法，被西方称为世界三大分类法；CC 虽然实际使用单位并不多，但作为典型的分面分类法，具有较大的理论意义和研究价值。

1. 《杜威十进分类法》

1) 发展概况

《杜威十进分类法》(DDC, Dewey Decimal Classification) 由美国著名的图书馆学家麦维尔·杜威(1851—1931) 创立，首次出版于 1876 年，取名为《图书馆图书、小册子排架和编目用分类法及主题索引》，是一本仅有 40 多页的小册子，收入近一千个类目，用三位阿拉伯数字作为号码。篇幅虽小，但因为它率先提出以相关排列法代替当时美国图书馆普遍采用的固定排列法，并首次为类目表编制相关索引，而且标记简明，所以受到了普遍欢迎。1885 年的第 2 版改名为《十进分类法与相关索引》，同第 1 版相比，第 2 版加深了类目的细分程度，首次为分类法的主表配置了标准复分表，并明确规定了在保持已有类目稳定的基础上进行修订的方针，为该分类法的进一步发展奠定了基础。自从 1951 年出版的第 15 版起，改称《杜威十进分类法》(以下称《杜威法》)，一直沿用至今。

从 1891 年第 4 版起，DDC 的编者就不是杜威本人了。1922 年，杜威委托普拉西湖俱乐部教育基金会组织一个常设委员会永远负责 DDC 的修订和出版工作。1958 年第 16 版，其修订工作实际移交给了美国国会图书馆继续进行。1975 年成立了十进分类法编辑政策委员会。目前，DDC 的出版工作由 OCLC 联机计算机图书馆中心的一个分支机构——森林出版社负责。

为了适应世界知识体系的不断演变，DDC 在诞生的一百多年里，一直积极地进行着更新和完善。

DDC 早在 20 世纪 80 年代就完成了计算机管理系统的研制，于 1993 年 1 月正式推出电子杜威(ED, Electronic Dewey) 第 1 版，这是世界上第一个自动化的交互式分类法系统。1996 年 7 月，森林出版社又在电子杜威的基础上，推出了基于 DDC 第 20 版的视窗杜威(DFW, Dewey for Windows)。同年，DDC 21 的视窗版又问世了。

视窗杜威(DFW) 对电子杜威(ED) 有所改进，操作平台由 DOS 升为 Windows，使用户界面有了根本性的变化。视窗杜威有单机版和网络版两种。用户可以通过类号、类名、术语等多种途径获取类目体系、索引及与类目有关的完整资料，并可同时调用多种不同功能的窗口进行查找、显示和处理。

2003 年 6 月，DDC 推出了第 22 版的电子版。同年 9 月，正式出版了第 22 版的印刷版，该版已由最初的 40 多页的小册子，扩展到如今的四卷、3000 多页的规模了。

DDC 是当今世界上影响最大、用户最多的图书馆分类法，目前共有 35 种语文版本，使用于 135 个国家的约 20 多万个图书馆，不仅用来组织图书馆藏书，而且也广泛用于书目和文摘数据库及网络信息资源的组织和检索。美国国会图书馆和英国不列颠图书馆发行的书目记录中均有 DDC 类号。

2) 体系结构

杜威编制 DDC 的指导思想是实用主义。他强调，实用就是一切，实用、方便是杜威编制 DDC 最重要的标准。

在类目安排上，杜威接受了美国哈里斯编制的图书分类法，只进行了少量调整。英国哲学家培根依据人的心理活动的功能区分知识，他认为人类的心理活动从低级到高级有三种功能，即记忆、想象和理性，依次产生出历史、文艺和哲学三大类知识。哈里斯分类法采用培根的知识划分原则，但对原有的次序进行了倒装，改为哲学、文艺和历史，将全部图书按照

当时美国高等学校设置的课程学科分为 100 类, 按照这个倒转了的培根体系加以排列, 就构成了哈里斯分类法的类目体系。

杜威则把哈里斯分类法的 100 类归纳为 10 个大类, 并基本上保持了原有的序列, 作为一级类目; 二级类目则与哈里斯分类法的 100 类基本相同。这样, 《杜威法》就间接地与培根的知识分类体系建立了密切的联系, 因此《杜威法》的分类体系被形象地称为倒转培根法。

在 DDC 中, 基本大类是按学科或研究领域组织的。DDC 最基本的原理是: 分类法按学科而不是按主题组织, 这也是等级列举式分类法的共同特征。DDC 的类目体系按照学科之间的一定联系与关系归纳组织, 但并非一门学科为一类, 不少相关学科成为类组。每一学科门类逐级按内容层层展开, 形成等级体系, 但并非机械地按 10 个类划分。类目体系采取对已知主题进行详细列举的方法, 在表内尽量为每一个已知的主题准备一个具体的位置。

DDC 的主表是对已知主题的详细列举, 基本上是以学科为中心展开的。其第一级类目首先将所有的学科门类分为 9 个大类, 再为不能归入任何一类的图书设一总类, 共 10 个基本大类。基本大类以后再设置 9 个二级类加一个总类。以此类推, 形成一个层层展开的十进分类体系。下面是 DDC 第 21 版的 10 个大类 (Classes)。

DDC 第 21 版的类目表共有 10 大类 (总纲)。

000 计算机、信息及总论 (Computers, Information & General Reference)

100 哲学和心理学 (Philosophy & Psychology)

200 宗教 (Religion)

300 社会科学 (Social Sciences)

400 语言 (Language)

500 自然科学 (Science)

600 技术 (Technology)

700 艺术和娱乐 (Arts & Recreation)

800 文学 (Literature)

900 历史和地理 (History & Geography)

每一大类之下又展开为 9 个类和 1 个“总论”类, 合成 10 个类, 称之为门 (Divisions)。

每一门之下, 通常又分为 9 个小类及 1 个“总论”性的类目, 而这些小类被称之为纲 (Sections)。

纲以下的类目也可按上述原则逐级细分, 细分出来的类目统称为子目。一般来说, 每一类也划分出 9 个小类和 1 个“总论”性类目。不足 9 类的留出空号, 多于 9 类的一般用扩九法, 即用其中一位数字配做“其他”类后, 再用其下一位小数展开。这样逐级细分, 形成一个详细列举的分类主表。子目以下可以进一步划分, 有的分到十几级, 直到不再有文献保证为止。

DDC 平均每 6 年修订一次, 它的修订原则有两条: 一是保持分类号连续性和完整性原则; 二是与知识发展保持同步原则。由于遵循这两条修订原则, 所以虽然历经一百多年的发展, 但 DDC 的体系结构变化不大。

3) 标记符号和标记制度

DDC 采用的标记符号是通俗易懂的阿拉伯数字 (也有少数类目可采用拉丁字母或其他符号作为标记符号的一部分)。全部数字符号按小数来理解, 按小数值的顺序来排列。为了醒目和便于阅读, 第三位与第四位小数之间用小圆点隔开。

DDC 利用小数的特点来标记其等级列举式类目体系, 分类号码按小数值的大小排列。

DDC 基本上采用小数层累标记制度, 用第一位小数标记第一级类目, 第二位小数标记

第二级类目，依此类推，层层隶属，层层包含。例如：

600	技术
620	工程学
621	机械工程
621.3	电力工程
621.38	电子与电力通信工程
621.384	无线电通信工程

当然，也有不少情况是小数的级位与类目所展开的层次不完全一一对应，小数的级位不能用来判断类目之间的上下位类和同位类关系，而只用来表明类目之间的前后顺序位置，采用的是小数顺序制。另外，有时也适当使用灵活的标记方法，用来扩充下位类。同时，重视号码的助记性，往往使用同一个数字表示相同概念。例如，在 DDC 标记中，理论总是 1、历史总是 9、欧洲总是 4、亚洲总是 5、中国总是 51 等，给分类工作者和用户使用类表带来了一定的便利。

4) 类目注释系统、复分表和类目索引

DDC 的类目注释系统丰富、详细，主要包括描述类目内容的注释、包括注释、涉及其他类目的注释、修订注释和标引方法注释。

由于 DDC 的历史沿革比较悠久，它的复分表是不断建立和发展起来的。在 17 版以前只有一个标准复分表，17 版增设了地区复分表，18 版以后又增设了 5 个通用复分表。目前，DDC 共有 7 个复分表，依次为：标准复分表，地理区域、历史时期、人物复分表，文学复分表，语言复分表，人种、种族、民族复分表，语种表，人员复分表。概括地说，这 7 个表可以划分为两类，即通用复分表和专类复分表。其中，文学复分表和语言复分表是专类复分表，其余 5 个均为通用复分表。

DDC 复分表的主要作用是精简类表篇幅，避免类目的重复列举，与此同时，增强了分类法的组配因素，使分类法的体系向列举与组配相结合的方向发展。

DDC 有一个详细的相关索引，它是杜威首创的一种分类法索引。该索引不仅收入类表中类名和注释中所有的术语及其同义词，而且还将一个主题的各个方面及被一主题所规定的词或倒装的词集中在同一标目之下，从而可以将分散在各个学科中同一主题对象的不同类目集中在一起，对于不理解学科之间关系的用户，可以通过主题字顺索引查找相关主题。

相关索引在 DDC 体系中是一个重要的部分，它包括类表和复分表中的大多数术语，以及类表和复分表所代表的概念中有文献保证的术语。

为了充分揭示词间关系，DDC 在索引款目之间使用参照的方法，用于指向上位词；同时，第 21 版和第 22 版还引入了处理多学科著作的新的机制。

在电子版杜威中，DDC 号码不仅与类名联系，而且还输入了《国会标题表》中对应的标题，用户可以通过类表、关键词和大类类号结合，查找对应的分类号、相关的分类号及对应的国会标题表中的标题，联机使用。在索引编制手段上，DDC 还利用 OCLC 联机目录中分类号与主题词的同现率作为编制依据，改进了索引的编制。

5) 主要贡献与不足

《杜威十进分类法》是当今世界上流传最广、影响最大的一部文献分类法，是现代文献分类法发展史上的一个里程碑。一百多年来，DDC 通过持续的修订和科学的管理，尽可能地与科学技术的发展保持同步，使类目体系能够不断容纳新主题、反映新观点，有意识地消除或淡化宗教及政治倾向性，使它有了更加广阔的发展空间，被越来越多的国家和地区用于藏书组织，目前已被翻译成几十种文字，其中包括全译本、节译本、增补本和改编本。

总结一下, DDC 对世界分类法的主要贡献同时也最值得其他分类法学习和借鉴的成就是:

- (1) 在文献排架和目录组织中首先使用了相关排列法;
- (2) 在标记制度方面首先采用了小数层累标记制;
- (3) 创建了等级分明的类目体系;
- (4) 首次配置了详细的相关索引;
- (5) 建立了对 DDC 进行定期修订的稳定的管理机构。

所有这些贡献和成就, 既使 DDC 建立分类检索系统和组织分类排架的实践职能有了可靠的技术支持, 又为作为世界著名分类法的 DDC 的未来的进一步完善和发展提供了强有力的组织保障。

当然, DDC 也存在一些明显的不足, 例如:

- (1) 由基本大类构成的类目体系对不同学科门类之间所具有的内在联系反映得不够充分, 影响了整个类目体系的学科系统性;
- (2) 大类的设置不能适应现代科学的发展, 不少已经过时的类目结构与现实文献的学科内容相脱离;
- (3) 过于突出美国中心的特征, 在使用和发展方面具有一定的局限性;
- (4) 小数层累标记制度使类号冗长, 不利于文献排架。

2. 《国际十进分类法》

1) 发展概况

《国际十进分类法》(UDC, Universal Decimal Classification), 也译做通用十进分类法, 是由比利时学者奥特勒和拉封丹共同编制的一部著名的列举—组配式分类法。该分类法也被称为世界上第一部半分面分类法, 所谓半分面分类法, 是因为体系结构主要是列举式, 而在同位类展开时需要进行相关的组配。它既不同于《杜威十进分类法》, 又有别于阮冈纳赞的《冒号分类法》。

1895 年, 奥特勒和拉封丹主持召开第一次国际目录学会议, 会上成立了国际目录学会, 决定编制世界书目, 需要有一个分类法作为世界书目的工具。当时 DDC 正出到第五版。奥特勒等决定以该分类法为基础进行改造。在获得杜威同意后, 1895 年开始对类表进行改造, 将类目加细, 并增设各种辅助符号供组配使用。第一版为法文版, 名为《世界图书总目手册》, 由国际目录学会于 1905—1907 年期间分 35 个分册出版, 共包括约 3 万 3 千个类目和一个字顺索引, 远比同期的 DDC 类表详细。该分类法得到了欧洲许多图书馆和文献单位的使用。第二版出版于 1927—1933 年, 仍为法文版, 类目增加到 7 万个, 改为目前使用的名称。为了强调分类法的国际性, UDC 的第三版改为德文版, 该版始于 1934 年, 其间因第二次世界大战而中断, 直到 1953 年才最后完成, 约 14 万个类目。此后的各种语文版本均是在法、德两个版本的基础上发展起来的。

目前 UDC 共有 23 种语文版本, 并有详本、中型本和简本的区别。一般情况下, 详本列表详尽, 大致在 20 万个类目左右; 中型本为详本的三分之一, 约 5~6 万个类目; 简本约为详本的十分之一, 即 2 万左右。其中, 详本往往按分册出版, 无严格的出版计划。为便于整个分类法的管理和发展, 1993 年建立了 UDC 的英文主文档, 将类目限定在 6 万个左右, 明确将其作为 UDC 各种语文不同类型版本修订和发展的基础。

UDC 是一种文献分类法, 其分类对象为各种类型文献, 包括小册子、科技报告和期刊论文等, 主要不是作为文献排架的工具编制的。目前, UDC 主要用于欧洲各国的专业图书馆、文献中心和情报机构, 不少文摘和索引工具也采用 UDC, 是国外使用较广的三大分类法之一。

2) 体系结构

UDC 的基本原理与 DDC 相同，突出实用这一编制思想，自称是一种实用的分类法。奥特勒明确提出：不应该将 UDC 看做是一种知识的哲学分类，类目的次序也不是最重要的。UDC 的目的在于：只要经过正确的标号和排序，就可以迅速地以任何角度查找到任何一篇文献。

UDC 是一个在详尽列举式类表的基础上大量采用组配方式建立的文献分类体系，由主表和辅助表及索引组成。

UDC 的主表是在 DDC 基本结构的基础上发展起来的一个层层展开的十进系统，也将人类知识分为 10 个大类（据 1989 年英文中型版）。

- 0 总类、科学和知识
- 4 （语言）
- 5 数学和自然科学
- 6 应用科学、医学、技术
- 7 艺术、娱乐、体育
- 8 语言、语言学、文学
- 9 地理、传记、历史

1964 年，第四大类的“语言”被并入第八大类“文学”内，空出来的第四大类的号码拟作为扩充科技类目之用，为分类法的进一步发展预留了空间，同时也进一步揭示了“语言”和“文学”在内容上的密切联系。

UDC 的大类序列与 DDC 基本相同，但大类标记由 DDC 的三位数字变为一位数字。

UDC 类目区分程度比 DDC 详细，按照从一般到特殊的原则，逐级进行区分，形成层层展开、详细列举的等级类目体系。例如：

- 6 应用科学、医学、技术
- 62 工程、技术（总论）
- 621 机械工程总论、核技术、电气工程、机械制造
- 621.3 电工程、电技术、电气工程
- 621.39 电信技术
- 621.396 无线电通信设备和方法
- 621.396.9 雷达
-

值得一提的是，UDC 的类目体系虽然是在 DDC 的大类体系的基础上展开的，但并没有完全照搬，而是组织了不同学科领域的主题专家对各个大类进行了独立的设置和展开，并且广泛应用了组配方式来表达复合主题，大量采用了分面结构，是名副其实的列举-组配式分类法。

3) 标记符号和标记制度

UDC 最突出的特点是：在详尽类表的基础上，结合辅助表和一系列复分标记的使用，广泛采用了组配的方式。

UDC 的标记由主表类号和各种辅助符号组成。

UDC 的主类号与 DDC 比较接近，采用单纯阿拉伯数字，以层累制方式配置号码，用个数（0~9）标记一级类，十位数（00~99）标记二级类，百位数（000~999）标记三级类，以下每扩展（细分）一级，就加一位数。为便于识别，每三位数字后加一小数点。

UDC 基本上采用小数层累制作为标记制度，分类号码的级位体现类目的等级。每细分

一次,就在原号码后加一位号码。当展开的同位类超过9个或需要预留空位时,采用八分法或百分法进行标记。

为了缩短号码和扩充类目,采用借号法(包括借上级号码、下级号码和同级号码)标记类目先后次序。

4) 辅助符号

UDC 不仅编有多种通用复分表与专类复分表,而且根据组配的需要设置了多种辅助符号。主表类号与各种辅助符号的组配,使复合主题得到了多方面的揭示,从而使分类法具有更大的灵活性。比较灵活的组配和复分是 UDC 一个很重要的特点。正是因为复分表和各种辅助符号的广泛使用,才使得 UDC 成为一种等级体系一分面组配式的分类法。

概括地说,UDC 共设有两种辅助符号:一是独立使用的通用辅助符号;二是结合辅助表使用的辅助符号。第二种辅助符号又可以细分为两种:结合通用辅助表使用的辅助符号、结合专门辅助表使用的辅助符号。

5) 修订和管理

20 世纪 80 年代前,UDC 的修订由国际文献联合会(FID)负责,由 FID 分类法中心委员会进行日常管理工作,FID 的成员国也建立相应国家的委员会,负责该种语文版本的管理。为了保持类表的稳定性,UDC 采用保留空类的方针,规定一个原有类号被修改放弃后,10 年内不得使用。由于修订过程缓慢,而且涉及不同语言和不同详略程度版本的差异,整个类表缺乏有效的整体控制,很难与知识发展保持同步,无法满足用户的实际使用需要。

为了有效进行类表的修订,20 世纪 80 年代以后,UDC 的修订和管理方式发生了很大变化,成立了 UDC 管理委员会,负责 UDC 的修订和管理。1992 年初,UDC 的管理权改为由 UDC 集团负责。UDC 集团的参加者,除 FID 以外,还包括比利时、日本、荷兰、西班牙、英国的 UDC 出版者,下设秘书处。同时,建立了以英文中型版为基础的核心主文档,该文档包括将近 6 万个类目,将作为整个 UDC 类表进一步发展的基础。

目前,UDC 的管理由专家组成的编辑小组实际负责,放弃了原有的修订方式,包括号码改变后 10 年内不得使用的方针,使修订工作更加迅速有效。同 DDC 一样,UDC 的修订分为两个层次,即常规修订和彻底修订。而且,努力促进类表逐步向分面发展,建立组配结构。

6) 主要优点与不足

UDC 比较突出的优点是:

(1) 在世界分类法发展史上首次将概念分析原理应用于文献分类标引实践,是组配分类的先驱;

(2) 类表列举详尽,组配灵活,既利用复分表进行组配,也大量使用各种辅助符号进行类目之间的组织,达到了充分揭示文献主题的目的;

(3) 标记符号表达性强,运用各种辅助符号表达文献的主题成分,可以轮排,以适应计算机检索的需要。

UDC 存在的一些不足是:

(1) 基本大类的设置缺乏均衡性;

(2) 组配规则过于灵活,影响了标引的一致性;

(3) 分类号码冗长,辅助符号繁多复杂,给手工排检带来了不便;

(4) 缺乏稳定而强有力的机构负责管理和修订工作,没有统一的修订方针。

3. 《美国国会图书馆图书分类法》

1) 发展概况

《美国国会图书馆图书分类法》(LCC, Library of Congress Classification)是为了适应美

国国会图书馆图书分类和排架的要求而编制的大型综合性分类法。

LCC 是由各个主题领域的专家，依据图书馆文献的特点，按照大类独立编制，分册陆续出版的。从 1902 年的“Z 目录学”大类首先出版，直到 20 世纪 80 年代整个分类法的各个分册才陆续出齐。整个分类法缺乏整体性，没有统一的合订本和统一的索引，有人称它是专业分类法的集合。

虽然 LCC 的编制初衷是为了美国国会图书馆藏书的排架和检索，但随着它的影响的扩大，使用范围越来越广，逐渐为许多大型学术性图书馆、研究图书馆及部分公共图书馆所使用。目前，LCC 已经完成了机读文档的建设。

2) 体系结构

LCC 的分类体系基本上是以学科为中心而建立的。21 个基本大类的设置参考了克特展开式分类法的体系，并以美国国会图书馆的收藏特点为依据。整个大类次序按照总类、哲学、历史和地理、社会科学、艺术和文学、科学技术的顺序组织。

- A 总类
- B 哲学、心理学
- C 历史：辅助科学
- D 历史：世界史
- E—F 历史：美洲史
- G 地理、人类学
- H 社会科学
- J 政治
- K 法律
- L 教育
- M 音乐
- N 美术
- P 语言、文学
- Q 科学
- R 医学
- S 农业
- T 技术
- U 军事科学
- V 海军
- Z 书目及图书馆学

大类以下，由各类专家根据学科领域的特点展开类目体系，一般先划分出基本学科或分支，然后再按主题、形式、地区、时代进一步划分，按照从总到分的次序逐级进行等级显示。

LCC 的详表是根据文献保证原则建立的，所以，各个基本大类的馆藏资源情况不同，就导致了各个大类之间类目展开的详略程度缺乏平衡性。

LCC 是典型的等级列举式分类法，对复合主题列举详尽。为了控制分类法的篇幅，LCC 也注意使用了为数不多的复分表，但与前面介绍的多数分类法中的复分表有所不同，主要表现在两个方面：一是没有通用复分表，只有专类复分表；二是复分表号码不是单独配置的，而是根据主表顺序号码中预先留下的空号确定的，也就是说，同一复分类目往往有多个号码，不具有助记性。LCC 采用的复分表类型主要包括：形式复分表、地区复分表、年代复分表、主题复分表、汇编复分表、著者复分表等。

3) 标记符号和标记制度

LCC 采用字母加阿拉伯数字的混合号码来标记类目, 类号通常由三部分组成: 一是大写字母, 通常以一位字母表示基本大类, 两位或三位字母表示其下位类; 二是阿拉伯数字 1~9999 表示其子类, 必要时可以用小数进行扩充; 三是在许多类下进一步用字母和数字组成克特号(书号), 再加上出版年代号。例如:

Z668 (分类号, 表示图书馆学目录学教学)

. w54 (克特著者号, 作者为 White)

1976 (出版年)

LCC 采用的是顺序标记制, 标记符号不能反映类目之间的等级关系, 缺乏表达性, 而且不利于在电子形式中通过标记符号对类表的等级加以显示。但是, 容纳性强, 也比较简短。另外, 前面讲过, LCC 的复分表的类号不具有助记性。

4) 修订和管理

美国国会图书馆编目方针和支持办公室主要负责 LCC 的修订工作, LCC 的编目人员起协助作用。

LCC 的各个类表没有统一的出版计划, 各表的新版或修订一般根据需要分别准备, 独立出版, 情况很不一致。

LCC 只有分册索引, 有的分册还没有索引。目前, 可供 LCC 使用的、以整个分类法为对象的书本式索引, 只有《国会图书馆分类法索引》和《国会图书馆分类法索引汇编》。这两本索引都是各个分册索引的汇编, 只能满足从主题字顺角度查找类目的基本需要。但是, 值得一提的是, 1995 年, LCC 整个类表已经被输入计算机, 可以说, 事实上, 已经建立起了 LCC 的完整的机读索引, 而且, 这个重要的结果必然会给整个类表的修订、管理和使用带来极大的便利。

5) 主要优点与不足

LCC 的优点主要是:

(1) 因为是一部依据文献保证原则编制的分类体系, 所以, 能较好地适应文献标引的需要;

(2) 类目体系由各学科专家编制, 适合研究性图书馆的分类特点;

(3) 标记简短, 容纳性强, 使用组配少, 便于号码配置;

(4) 以日常编目工作为修订依据, 增补和变动比较及时;

(5) 类表结构稳定, 类目体系变动较少, 有利于实际编目工作。

LCC 存在的不足主要是:

(1) 分类法的编制及修订工作均缺乏明确的理论指导, 降低了类表应有的系统性和规律性;

(2) 因受组织和检索美国国会图书馆藏书的制约, 削弱了整个分类法的通用性, 并且以西方为中心的倾向也比较明显;

(3) 类表从一开始就按照分册编制、修订, 缺乏整体性;

(4) 详尽列举方式使类表篇幅巨大, 增加了类表管理和更新的费用;

(5) 采用的顺序标记制虽然适用于藏书排架, 但表达性差, 尤其是在计算机系统中不利于通过标记对类表进行等级显示。

4. 《冒号分类法》

1) 发展概况

《冒号分类法》(CC, Colon Classification) 是印度著名图书馆学家阮冈纳赞所创制的分

面分类法。

阮冈纳赞最初编制的分类法，还是受到了一种名为“梅卡洛”（Meccano）的机械组合玩具的启发。这种玩具的全部材料只有槽条、齿轮、螺钉、螺帽等 12 个小型零件，通过对它们进行适当的变换和组合，就可以装配成各种不同样式的玩具。阮冈纳赞由此联想到了图书分类表，设想着只用少而短的分类表，随时根据类表中各种成分适当加以变换，组配成任意主题类号。有了这一大胆的设想，阮冈纳赞随即投入了全部的热情和努力，开始了具有非凡意义的《冒号分类法》的编制工作。1925 年，在从英国返回印度的旅途中，阮冈纳赞终于完成了 CC 的初稿。随后，又经过了几年的实践，《冒号分类法》的第 1 版于 1933 年正式面世。

阮冈纳赞提出，为了随时反映和处理不断产生的新主题，文献分类表必须具有随时扩充的能力，具体来说，文献分类表必须具有这四个方面的无限容纳性：对于从属类目，要能够实现无限地细分；对于同位类目，要能够实现无限地增加；在同一类目的各个属项之间，要能够实现无限地相互组配，进而产生新的类目；在不同类目之间，要能够实现无限地相互联系，进而形成新的类目。

阮冈纳赞认为，文献分类法要实现上述功能，最关键的问题就是标记制度。

以往分类法的各种标记都有一个不可克服的缺点，即特定主题的分类号是固定的，这种先组式的类号必然使类目体系受到了很大的限制。

正是为了克服分类法的这个缺点，阮冈纳赞提出了面的分析法和分面标记法。面的分析法是分析概念内容的方法，也就是主题分析的一种方法。这种灵活的、不固定的、后组的、不先组的标记方法，基本具有阮冈纳赞所提出的四个方面的无限容纳性。

对于文献分类与知识体系的关系，阮冈纳赞的观点十分明确。他认为，文献分类应当以文献中包含的知识的类别为依据。人类的全部知识可以划分为若干个大类和惯用类，大类和惯用类都是基本类，而文献中的主题又总是由某个基本类包含的概念所构成的，因此，将每个基本类划分成许许多多的概念，通过对这些概念的相互组配，就可以生成许许多多的主题，而文献就可以根据这些主题进行归类。

到目前为止，CC 已经出了 7 版：1933 年正式出版第 1 版，1939 年出版第 2 版，1950 年出版第 3 版，1952 年出版第 4 版，1957 年出版第 5 版，1960 年将第 1 卷修订出版作为第 6 版，1972 年出版第 7 版。

CC 与以往文献分类法最大的不同之处就在于：充分肯定了分类语言具有动态性的特征。阮冈纳赞认为，知识是“多维的”、“动态的”、“无限领域的”。有了这种对知识问题的见解，便确定了 CC 适应知识领域动态发展的基调。为了能够使分类法最大限度地与知识发展同步，阮冈纳赞提出了具体的解决方法——分析与合成的方法，即通过对事物进行“基本范畴分析、相分析、面分析、巡分析、层分析”这种分析—合成程序，来完成对任何一种文献主题的描述。

阮冈纳赞以概念分析与综合原理编制的 CC，标志着分类语言的一个新发展，对当代分类法的理论与实践产生了广泛影响，目前，人们已经开始探讨将 CC 及其分面分析的原理运用于网上信息组织。

2) 类表结构

CC 的类表结构主要包括基本大类、惯用类、基本类、分面分析法等内容。

如前所述，分面分类法与体系分类法一样，也编列有基本大类。在基本大类的设置方面，阮冈纳赞认为，基本大类的数量不是固定不变的，可以随着新学科门类的出现而随时增加，大类的排列次序也并不十分重要。

CC 基本大类的设置,基本上是以传统学科门类为基础的,根据习惯上的传统学科门类划分出一定数量的基本大类,而且,每出一新的版本,基本大类的数量都有所改变。例如,第6版的基本大类如表3.3所示。

表 3.3 冒号分类法基本大类及其标记符号

标 记 符 号	基 本 大 类	标 记 符 号	基 本 大 类
z	综合类	LZ	药学
1	知识全体	M	实用技艺
2	图书馆学	△	精神体验与神秘主义
3	图书学	MZ	人文科学与社会科学
4	新闻学	MZA	人文科学
A	自然科学	N	美术
AZ	数理科学	NZ	文学与语言
B	数学	O	文学
BZ	物理科学	P	语言学
C	物理学	Q	宗教
D	工程学	R	哲学
E	化学	S	心理学
F	工业技术	Σ	社会科学
G	生物学	T	教育
H	地质学	U	地理
HZ	采矿学	V	历史
I	植物学	W	政治学
J	农业	X	经济学
K	动物学	Y	社会学
KZ	畜牧学	YZ	社会工作
L	医学	Z	法律

解说性大类列举:

- | | | | |
|-------|--------|-----|------|
| (: g) | 评论方法 | (P) | 信息论 |
| (p) | 会议方法 | (X) | 管理方法 |
| (r) | 行政报告方法 | | |

以上基本大类共划分为4个分区:z综合类为第一分区类目,以小写拉丁字母表示;以阿拉伯数字表示为其他学科所不能包括的新兴学科,作为第二分区类目;大写拉丁字母A~Z、△、Σ表示传统学科类目,作为第三分区类目;第四分区为新出现的各种方法学、带有工具性质的类目,以带有“()”的大写或小写拉丁字母表示。

所谓惯用类,就是根据传统习惯予以区分的一系列类目。大类和惯用类均称为基本类。基本类及其标记符号如表3.4所示。

表 3.4 基本类及其标记符号

标 记 符 号	基 本 类	标 记 符 号	基 本 类
H	地质学	R	哲学
H1	矿物学	R1	逻辑学

续表

标 记 符 号	基 本 类	标 记 符 号	基 本 类
H2	岩石学	R2	认识论
H3	构造地质学	R3	形而上学
H4	动力地质学	R4	伦理学
H5	地层学	R5	美学
H6	古生物学	R6	优惠系统（1）
H7	经济地质学	在此展开印度哲学	
H8	宇宙的假说

基本类是图书分类的出发点，只有基本类才能附有分面公式。

阮冈纳赞把概括文献主题的分类特征归纳为 5 种基本范畴，即本体（Personality）、物质（Matter）、能量（Energy）、空间（Space）、时间（Time）。

本体，指事物本身的各种体现；物质，指构成事物的材料；能量，指事物的各种活动、影响、状态和问题；空间，指事物存在或发生的地点；时间，指事物存在或发生的时期。

5 个基本范畴就是 5 个基本分面。阮冈纳赞以 P、M、E、S、T 分别代表这 5 种基本范畴（基本分面），用逗号“，”、分号“；”、冒号“：”、句点“·”、倒逗号“‘”5 种分面符号依次表示这五种基本范畴（基本分面）。

分面公式（Facet Formula），即分面的组配次序与标记符号，也称为组面公式。各个分面依据具体性递减的原则确定先后次序。

CC 总的分面公式为：“，[P]；[M]：[E]·[S]‘[T]”。以总的分面公式为依据，各类又有具体的分面公式。例如，“S 心理学”类的分面公式为 S[P]：[E][2P]。

巡（Round），是指在同一主题内数次出现的某种分面（能量、本体、物质分面等），如 [2E] 称为第二巡能量，[3E] 称为第三巡能量。

层（Level），是指同一巡内同范畴数个分面，如 [P2]、[P3] 分别称为第一巡第二层本体、第一巡第三层本体。

巡表示同一主题的纵向深入，而层则表示同时采用多种分类标准的横向联合。

3) 主要优点与不足

CC 是一部完全按照分面方式编制的分类法，与传统体系分类法相比，它突出的优点是：

（1）类表十分简练，标记的表达性强，可以在确切揭示文献主题的同时充分揭示复杂主题的关系；

（2）类表采用分面组配结构，对新出现的主题具有比较强的接纳能力和揭示能力，能较好地适应科学技术的变化发展，成为一种不同于传统分类法的全新的分类法类型。

CC 对分类法理论和技术的发展具有革命性的贡献，所提出的分析兼综合的原则，分面分析、分面标记的学说，极大地丰富了分类法的理论，对世界范围内分类法的编制和修订都产生了巨大影响，并且深刻地影响着整个知识组织理论，在计算机广泛应用的今天，具有巨大的理论价值和实用价值。

当然，CC 也存在一些不足之处：

（1）大类结构以神秘主义为中心展开，对整个分类体系依据的思想没有明确的说明；

（2）标记结果虽然表达性强，但标记方法复杂、符号种类繁多，给类表的使用造成困难；

（3）类表的展开不够均衡，有的大类采用深度分类，有的大类仅仅是概略分类；

（4）虽然在理论上代表着分类法发展的新的阶段，但类表本身的编制水平尚有待于提高，编辑和印刷错误较多，影响使用质量。

5. 《中国图书馆分类法》

1) 编制

《中国图书馆分类法》(CLC, Chinese Library Classification, 以下简称《中图法》)。前3版的名称均为《中国图书馆图书分类法》, 由于分类法使用的范围不断扩大, 第4版改为现在的名称。

《中图法》是在我国文物事业管理局的支持下, 由北京图书馆倡议, 集中全国36个大型文献单位的力量共同编制的一部大型综合性文献分类法。该分类法的编制广泛汲取了建国后文献分类法成功经验, 其中尤与《中小型图书馆图书分类法草案》(简称《中小型表》)和第一次编制的《中国图书馆分类法》(简称《大型法》)有着深厚的渊源。

1973年发行了《中图法》试用本, 在全国范围的图书馆中广泛征求意见后, 进行了认真的修订, 1975年《中国图书馆分类法》第一版正式出版, 并陆续为全国图书馆和文献单位所使用。其后, 除了根据本身使用的需要定期修订外, 还进行了不同版本和配套产品的编制, 逐步形成了以《中图法》为中心的版本系列, 1999年第4版出版, 改名为《中国图书馆分类法》。2010年9月出版第五版。

2) 类目体系

《中图法》根据毛泽东关于知识分类的思想, 将人类全部知识划分为哲学、社会科学、自然科学三个部分, 并以此作为确定分类法基本结构的理论依据。除此以外, 认为马克思主义、列宁主义、毛泽东思想是分类法编制的指导思想, 故将其作为特殊部类列于首位。而由于有些图书内容庞杂, 类无专属, 无法按某一学科内容性质分类, 所以将其概括为“综合类图书”, 也作为一个部类, 置于最后。

这样, 《中图法》便确定了五个基本部类: 马克思主义、列宁主义、毛泽东思想, 哲学, 社会科学, 自然科学, 综合性图书。

除了第一个和第五个基本部类以外, 其他三个基本部类因为是对知识的科学划分, 所以它们的序列反映了知识的内在联系。

五大基本部类只对人类全部知识进行了粗略的划分。《中图法》在照顾各学科领域平衡的基础上, 以国际上通用的基本学科划分和专业划分为依据, 同时考虑习惯的知识领域划分, 在五大基本部类的基础上, 进一步将社会科学部类扩充为9个基本大类, 将自然科学部类扩充为10个基本大类, 共设置了22个基本大类。

《中图法》第1版与第4版基本部类与基本大类对照列表如表3.5所示。

表 3.5 《中图法》第1版与第5版基本部类与基本大类对照表

第1版	第5版
A 马克思主义、列宁主义、毛泽东思想	A 马克思主义、列宁主义、毛泽东思想、邓小平理论
B 哲学	B 哲学、宗教
C 社会科学总论	C 社会科学总论
D 政治、法律	D 政治、法律
E 军事	E 军事
F 经济	F 经济
G 文化、科学、教育、体育	G 文化、科学、教育、体育
H 语言、文字	H 语言、文字
I 文学	I 文学
J 艺术	J 艺术

续表

第1版	第5版
K 历史、地理	K 历史、地理
N 自然科学总论	N 自然科学总论
O 数理科学和化学	O 数理科学和化学
P 天文学、地球科学	P 天文学、地球科学
Q 生物科学	Q 生物科学
R 医药、卫生	R 医药、卫生
S 农业科学	S 农业科学
T 工业技术	T 工业技术
U 交通运输	U 交通运输
V 航天、航空	V 航空、航天
X 环境科学	X 环境科学、安全科学
Z 综合性图书	Z 综合性图书

对于22个基本大类的设置及其序列,《中图法》在“第1版编制说明”中进行了详细的阐述,重点强调了所包含各学科之间的内在联系。

可以看出,社会科学领域中,按照政治、经济、文化的次序,在社会科学总论之后,首先列出政治、法律大类及与政治有密切联系的军事大类,其后序列经济类,然后再排列有关文化事业的类目,以及语言、文学、艺术等,最后为系统研究和阐述人类社会过程的历史科学及其密切相关的人文地理。自然科学部分中,类表按照基础科学和应用科学分别设类。基础科学遵循从简单到复杂、从低级到高级的次序排列。应用科学的排列中,首先列出与生物科学有密切联系的医药、卫生和农业科学,其后再序列工业技术门类,其中,交通运输、航空航天根据学科发展分别设为基本大类,列于工业技术之后;并将环境科学、安全科学作为保护人类生态环境、维护人体安全的综合性科学设为一个独立大类。

在基本大类的基础上,《中图法》根据各类文献的特点,遵循从总到分、从一般到具体、从理论到实践的方式逐级展开。根据哲学、社会科学内容及其发展与国家、时代具有密切联系的特点,哲学、社会科学领域各类在多数学科的理论方法之后,对涉及各国情况的类目,一般先按国家序列,然后再按其他标准分类,时代的区分则通常在国家之后根据需要进行。自然科学各类中,基础科学一般按研究对象的性质,依据从总到分、从简单到复杂的次序序列;技术科学和应用科学类目的编列通常先列出基本理论或技术科学,然后再按对象列出各具体部门,并按先总后分的次序设类。对自然科学各类中国家与时代的区分,仅限于学科史、现状等有关的少数类目。

为了使文献分类法适应现代文献分类的特点及我国文献分类的实际需要,《中图法》还在类目设置中采用了多种处理方法。

(1) 在自然科学的学科门类中,对科学技术的新成果、新技术,根据需要予以充分反映,在不影响科学性的前提下,适当突出它的级位。例如,“地球科学”中的“地质力学”、“海洋学”,“生物科学”中的“分子生物学”等,都被编列在较为显著的位置,以适应新学科的发展。

(2) 为了满足特定学科专业文献组织的需要,在部分特定的学科知识门类下集中设类。例如,部分放弃了严格按照学科列类的方式,将工业技术中的“制药化学工业”下的有些类目设置在“医药、卫生”大类之中,使“医药、卫生”大类的类目相对集中,能够比较充分

地满足医学文献单位的实际需要。

(3) 在处理交叉关系类目时, 突出其规律性、灵活性和适用性。对于多重相关、多重隶属关系的类目, 除了按照研究对象所属的学科归类外, 一般还揭示其相关联系, 为用户提供各种选择的可能, 具体做法包括: 编制交替类目, 编制类目参照, 规定互见分类方法, 通过注释指明集中和分散的方法等。

(4) 根据文献处理的需要多重列类。例如, 对于“TN949.1/.299 各种电视”类, 类表同时采用“按体制分”和“按功能、用途分”两种标准进一步区分类列, 并在类下明确规定处理办法: “涉及多重列类标准的著作入最后编列的类”, 使类表具有多角度处理文献的能力, 并具有分面组配的潜在能力。

(5) 双表列类。《中图法》在“D9 法律”类采用了双表列类的方法, 即为法律类编制了两个分类体系, 以满足不同文献单位对法律文献组织的不同需要。具体来说, 第一个类表采用先国家后法律部门的顺序展开类目体系, 可满足一般文献单位的使用需要; 第二个类表采用先法律部门后国家的顺序展开类目体系, 可满足法律专业文献单位的使用需要。这样, 增强了类表的灵活性和适用性。

《中图法》的复分表包括两种类型: 通用复分表和专类复分表。

《中图法》有 8 个通用复分表, 列于主表之后, 依次为: 总论复分表; 世界地区表; 中国地区表; 世界时代表; 中国时代表; 世界种族与民族表; 中国民族表; 通用时间、地点表。

在编制手检工具时, 通用复分表只对主表类目起复分作用, 不能单独使用。

在用于机读数据标引时, 利用各通用复分表的子目号都有特定的前置符号区分的特点, 可以将它们均作为一个独立的检索标识从主类号中剥离出来, 成为机读数据中能参与组配检索的标识。换言之, 凡是一个类目没有按照其他标准复分、仿分的注释时, 文献主题中的地区、时代、民族、种族、通用时间地点要素, 都可以把相关的通用复分子目号单独著录于一个 690 字段。在机读数据 690 字段标引时, 按照规定, 总论复分号在某些情况下也可以单独著录于一个 690 字段, 而且可以重复使用。

《中图法》除了 8 个通用复分表外, 在主表中的有关类目下还编有 58 个专类复分表, 供相关类目细分时组配使用。

仿分及使用关联符号进行类间组配等方法, 在《中图法》中也得到了广泛的应用。

各类复分表及各种复分方法的应用, 对于《中图法》的编制和使用所起到的最突出的作用是: 大大缩短了类表的篇幅, 加强了类表对主题的揭示能力, 增强了类表标引和检索的灵活性和规律性。

3) 标记符号和配号制度

《中图法》采用汉语拼音与阿拉伯数字相结合的混合号码, 一般以一个大写字母标志一个大类。在工业技术大类中, 为了适应工业部门分类的需要, 采用双字母方式标记二级类目。其余类目均采用数字标记, 所有数字按小数对待。为了使号码醒目, 规定每三位数字加一圆点间隔, 圆点不包含任何意义。

《中图法》的配号制度基本上遵循层累标记制的原则。数字部分使用小数制编号, 即首先顺序字母后的第一位数字, 然后顺序第二位, 以此类推。分类号的排列严格按照小数制的排列方法。

数字设置时, 尽可能地使号码的级数与类目的级数相符。

但是, 为了满足标记对类目容纳性、简短性的要求, 在号码配置上通常采用各种灵活的标记方法, 包括八分法、双位制、借号法、预留空号法、字母标记法、对应编号法等。

另外, 《中图法》还采用了如下一些辅助符号。

(1) “a”，推荐号，供马列主义经典作家的著作在各相关类目中重复反映时使用，排列于相同号码之前。

(2) “—”，总论复分号，为总论复分表号码的组成部分，带总论复分号的标记排列于相同的主类号之后、下一个分类号之前。

(3) “:”，关联符号，用于两个类目之间的组配，通常依据主表中有关类目下的注释，在按照某一类目下相对集中的资料进一步排序时使用。

(4) “()”，国家区分号，通常在分类法未要求区分国家的情况下，供收藏文献较多的单位扩大地区复分表的应用范围时使用。

(5) “=”，时代区分号，通常在分类法未要求区分时代的情况下，供收藏文献较多的单位扩大时代复分表的应用范围时使用。

(6) “ ””，民族、种族区分号，在分类法未要求区分民族的情况下，供收藏文献较多的单位扩大民族区分表的应用范围时使用。

(7) “<>”，通用时间、地点区分号，用于区分国家、地区、时代以外的空间和时间概念。

(8) “+”，联合符号，用于联结并列关系的主题。

《中图法》规定，在同一个类目之下，以上各种辅助符号的排列次序是：—，()，“ ”，=，<>，:，+。

4) 修订与管理

第一次修订。由于《中图法》是在特殊时期编制成的，当时受“左倾”思潮的影响很深，类表中有不少不科学、不合理的类目，随着科学技术的发展，人们思想观念的改变，亟须对其进行必要的修改和补充。根据 1979 年长沙会议确定的《中图法》修订方针和原则，开始了对《中图法》第 1 版的修订工作。1980 年修订工作完成，正式出版《中图法》第 2 版。

第 2 版《中图法》重点对第 1 版中的极“左”思潮所产生的影响加以清除，如改变了按政治观点列类的方法；改变了教育类中先按国家划分的类目体系；合理地安排了建国前与中国有关的类目。另外，自第 1 版出版以来，科学技术迅速发展，与此相关的文献内容、数量、种类也都发生了巨大的变化，显然，已有的类目体系不利于对当时的知识整体、文献整体的全面揭示，为此，《中图法》第 2 版增补了管理学、系统学、遥感技术、遗传工程等重要学科类目。

《中图法》第 2 版修订增加了 1173 个新类目，删除了 870 个类目，改动了 577 个类目。也对有些类目注释、复分表进行了一定的修订。

第二次修订。总体来看，1990 年出版的《中图法》第 3 版，对第 2 版的各级类目及标记系统进行了全面的检查、调整和扩充，进一步清除了“左倾”思潮对分类法的影响；把“安全科学”与“环境科学”合并为一个类组；使用双表列类法建立了法律类第二分类体系；扩大冒号组配的使用范围；等等。

《中图法》第 3 版修订的指导思想是：“分类法是按照学科系统编制的，其体系结构和类目设置，必然要受一定时间条件的局限。为此，随着科学技术的发展而不断加以充实、修订，是必然的。但它毕竟是类分文献的工具，它的编制和使用具有连续性和稳定性，应尽量避免大的更动。”

在此思想的指导下，《中图法》第 3 版解决了第 2 版的一些遗留问题。无论是体系结构还是类目设置均比第 2 版更为科学、合理；在编制技术上，第 3 版采用了新的技术，比第 2 版有进一步的改进与提高。

《中图法》第 3 版由于类目的增补、删除、改动而使相关图书必须随之进行改编的地方

多达 2190 多处。

第三次修订。1999 年出版的《中图法》第 4 版在修订方针上,坚持基本部类和基本大类的设置和序列基本保持不变;强调编制使用的连续性、稳定性,对标记的变动持谨慎态度,字母—数字混合的标记符号与层累小数制的标记制度基本保持不变。在这个前提下,实现个别大类类目体系结构的调整与完善,增加与补充新学科、新事物主题概念。同时,规定各个版本之间保持基本体系结构一致,将其区别限制在类目划分深度、组配方法的使用上;充分利用并不断完善已经采用的分类方法和技术,如多重列类法,交替类目法,参见类目法,类目复分、仿分法,主类号直接组配法等。

在这个修订原则的指导下,《中图法》第 4 版明确了类目体系调整的范围和重点。

第 4 版重点修订了“F 经济”、“TN 无线电电子学、电信技术”及“TP 自动化技术、计算技术”三个大类。

《中图法》对这三大类的修订主要采取了增补、扩充类目的方法。

除了将上述三大类作为重点修订以外,《中图法》第 4 版还全方位地对整个分类法的类目体系进行了调整,扩充或加细了许多类目,增加或改动了大量类目注释。对复分表也做了大幅度的增补和扩充。

此外,还通过增设指示性类目、增加沿革注释、修改或规范类目名称、规范注释用语和注释引用次序及扩大采用编列交替分类体系的列类法等方法,尽可能地使类目体系与知识体系、文献体系之间建立比较密切的关系,使分类法能够准确、合理地类分文献。

在管理机构上,《中图法》目前采用二级管理方式。

第一级,《中图法》编委会。其成员由参加编制分类法单位的代表组成,负责确定类表发展方针和全面修订,并在北京图书馆设置分类词表组,其职能是集中用户意见,负责类表日常修订工作。

第二级,专业分类委员会。负责该学科类表的修订组织工作和以《中图法》为基础的专业分类法的编制工作。

在管理方式上,由北京图书馆词表组负责对类表的日常维护,通过定期出版《〈中图法〉与文献标引信息简报》公布修订信息,与用户沟通,以便使类表能够适应实践发展的需要;全面修订则由编委会组织各科专家定期进行,一般是 6~8 年修订一次。

1981 年国家标准总局转发了“关于《中图法》作为国家试行标准草案的建议”,目前,《中图法》已经被国内各类型图书馆与其他文献信息机构广泛应用。据统计,大约有 90% 以上的图书情报单位使用《中图法》,我国集中编目部门均将《中图法》作为主要分类标引依据。

1985 年《中图法》获得国家科学技术进步一等奖,成为我国图书情报界唯一获得最高奖项的研究成果。

2001 年《中图法》的电子版正式出版,目前设置有专用网站: <http://clc.nlc.gov.cn>。目前,《中图法》的版权归国家图书馆所有。

5) 主要优点与不足

概括地说,《中图法》具有以下优点。

(1) 基本大类设置比较合理。全表以学科为中心建立分类体系,大类设置数量合理,基本反映了现代科学发展的现状和文献状况。在类目排列上,基本以类目内容之间的联系作为类目设置的依据,符合现代文献分类的特点。

(2) 类目体系的展开比较系统、适用。整个类表按照从总到分、从一般到个别、从低级到高级、从理论到实践的顺序展开,体现出较强的规律性。为了兼顾文献组织与文献检索的

双重功能，类目展开的级次一般控制在6级左右。

(3) 重视类表的灵活性。类表在保持类目体系展开的规律性的同时，广泛采用各种技术手段，为交叉学科、边缘学科等提供了各种选择的可能，如设置交替类目、类目参照、双表列类等，使用户在使用类表时有一定的选择余地。

(4) 追求编号技术的最佳结合。类表采用字母、数字结合的混合号码，以层累标记制方式配号，类号简明，同时使用一些特殊标记技术如八分法、双位制等，增强类目体系的容纳性。为了加强类号的助记性，设置了多种辅助符号供组配时使用。

(5) 发展了适合各种规模和对象的文献标引与检索的配套产品。包括根据不同文献规模、对象而编制的分类表、专业类表及《中国分类主题词表》等，为我国文献标引和检索的一致和规范提供了必要条件。

(6) 管理健全、规范。建立常设机构负责类表的修订和管理，并且有明确的管理形式和修订方针，保证了分类法能与知识发展同步，不断满足用户的分类标引需要。

同时，应当指出，《中图法》目前存在如下的主要问题。

(1) 部分类目体系需要进一步完善，如“G文化、科学、教育、体育”类，显得缺乏学术性、系统性。

(2) 通用复分表还不够充分、集中。例如，与国外一些著名分类法相比，《中图法》至今还没有编制语种、人物等复分表，影响了对一些主题进行充分的分类标引；地区表分为《中国地区表》和《世界地区表》，时代表分为《中国时代表》和《国际时代表》，使类目设置分散等。

(3) 分类法系列中不同类表之间需要加强协调。例如，从实际应用的情况来看，迫切需要对《资料法》进行适当的分面改造，以增强其对文献主题的组配标引能力；对于综合性分类法与专业分类法之间的关系处理，所需做出的相对具体、明确的规定等，都必须在实践的基础上逐渐解决。

《中国图书馆分类法》（第5版）在《中图法》（第4版）的基础上对类表进行了较大幅度的增、删、改：对D、G、T、S大类进行了局部调整；增补了大量新主题类目；合并使用频率过低的类目；删除重复或列类不当的类目；修改类名，增强类目的容纳性；增改注释，控制划分深度；增加了复分标记和“一般性问题”的禁用标记；完善了类目参见注释；修改补充了类目反向参照注释；增补了“通用时间、地点和环境、人员”附表的复分类型。此次修订新增类目1630多个，停用和删除类目约2500多个，修改类目约5200多个。例如，合并G25、G35图书馆学情报学体系。修改G25类名为“图书馆事业、信息事业”，G350/359的全部类目体系合并到G25/259类目体系中，同时增设同位类G254.9信息检索。

3.3 主题语言

3.3.1 主题法概述

1. 主题法的含义

“主题”一词，在不同的语境中有不同的解释。在信息组织中主要指信息资源所论述的主要对象，包括事物、问题、现象等。那些经过选择，用来表达信息资源主题的语词，则称为“主题词”。“主题法”是指直接以表达主题内容的语词作为检索标识，以字顺为主要检索途径，以参照系统等方法揭示词间关系的标引和检索信息资源的方法。主题法实际上包含两个含义：第一，是指信息资源的主题整序方法，即用语词标识标引信息资源和组织检索系统的方法；第二，是指主题语言。也就是说，主题法包含主题标引和主题语言。主题标引是指

对信息进行主题分析,用主题语言表达分析出的主题,赋予信息资源主题标识的过程;而主题语言是一种检索语言,标题词、元词、叙词等主题词就是主题语言的主体。

2. 主题法的原理

(1) 直接以语词作为主题标识。

主题法不像分类法那样,以一种抽象的号码系统作为分类标识,而是直接选用自然语言中的语词作为主题标识。例如,“土壤生态学”这一主题,在《中图法》中的分类标识为 S154.1;但在主题法中,可直接用“土壤生态学”作为主题标识,比分类标识直观。

(2) 以字顺序列作为主要检索途径。

虽然主题法往往也采用范畴(分类)、词族(等级)等方式对主题词进行组织,但字顺方式始终是它的主要排检依据。我国的主题检索系统通常是根据汉语特点、按照拼音或笔画笔顺进行排检的,因此,在使用主题法检索时,只要知道检索对象的名称,就可以按相应的字顺排检方式进行查找,方便快捷。在采用机检系统的情况下,一般可以直接输入语词,由计算机进行查找,不必像使用分类法那样必须预先了解主题所属的学科,因此通用性较好。

(3) 以主题为中心集中信息资源。

分类法由于受学科体系的限制,从不同学科角度研究同一对象的信息资源是被分散在各知识门类的,主题法则没有这一限制,而是直接从主题对象的角度揭示信息资源,这一特性是由主题标识和字顺排列决定的。以论述葡萄的文献为例,在分类法中,关于葡萄的栽植、葡萄的酿制、葡萄的贸易等主题,一般应按学科分别归入农业科学、工业技术、经济等不同学科门类,而在主题法中,通过语词标识和字顺排列,这些有关葡萄的各种信息资源可以直接在“葡萄”这一主题标识下集中予以揭示。

(4) 通过参照系统等方式揭示主题词之间的关系。

为了在采用字顺序列的同时有效揭示主题概念之间的关系,主题法发展了完备的参照系统,通过在主题词下设置用、代、属、分、参等多种参照项,建立起“隐蔽的分类体系”。此外,一些还备有词族索引、范畴索引、轮排索引等这些辅助索引,从各种不同的角度主题之间的关系。通过上述各种形式的结合,在主题词之间建立起充分的语义联系。当然,各种主题系统中对词间关系的揭示状况是不平衡的,就整体而言,其主题之间关系的揭示不如分类法。

主题法通常用于建立各种检索工具,不仅用来编制各类手工检索的书目索引、主题索引,同时也广泛用于组织机检系统,供计算机检索使用。与分类法相比,主题法的特点是可以集中与一个主题有关的各个方面的信息资源,检索的直接性好,适合进行各种特性检索,在性能上具有与分类法相互补充的特点。所以,主题法是从主题内容着手对信息资源进行组织的方法,但有其自身的规律和特点,是一种和分类法不同的信息组织和检索的方法。

3. 主题法的类型

主题法的类型可以有许多不同的划分,按照选词方法,可以分为标题法、元词法、叙词法、关键词法等;按照主题词使用时组配的先后顺序,可以分为先组定组式主题法、后组式主题法和先组散组式主题法;按照使用时是否对主题词进行控制,可以分为受控主题法与非受控主题法。

1) 按选词方法划分

(1) 标题法。

标题法是一种以标题词作为主题标识、以词表预先确定的组配方式标引和检索信息资源的主题法。所谓标题词,亦称标题,是指经过词汇控制,用来标引信息资源主题的词或词组,

通常为比较定型的名词术语。例如“电子计算机”、“信息资源”、“主题法”、“教学理论”等都可以作为标题词。目前,世界上使用最广泛的标题词表是《美国国会图书馆标题表》。

标题法除了直接选取自然语言中的单词和词组作为标题标识外,还采用其他一些标题形式。如复分标题(多级标题):肿瘤—治疗;肿瘤—治疗—中药;水果—病虫害;等等。倒置标题:飞机,低速;贸易,多边;等等。带限义词的标题:运动(哲学);运动(体育);等等。复分标题可对一个主题的各个方面的特征进行专指标引,倒置标题可集中一个主题的相关信息资源,带限义词的标题可明确标题词的专业含义。

标题法主要通过设置参照的方式对标题词之间的联系进行揭示。标题词之间的关系主要有三种:等同关系、等级关系和相关关系。等同关系揭示标题词与非标题词之间的关系;等级关系和相关关系都揭示标题词之间的关系。

等同关系是指一标题词与其含义相同,可以互相代替的语词之间的关系。对于等同关系的揭示,标题法所采用的参照符号是“见和见自”(See and See from)，“见”参照符号后指出的是标题词，“见自”参照符号后指出的是非标题词，如：

电脑	电子计算机
见 电子计算机	见自 电脑
形势几何学	拓扑学
见 拓扑学	见自 形势几何学

在标题法中,只能用标题词标引和检索信息资源,非标题词不能用来标引和检索信息资源,只能作为检索入口,但仍保留在标题表中。

等级关系是指广义词(泛指词)与狭义词(专指词)之间的关系,广义词是上位词即上级标题词,狭义词是下位词即下级标题词。揭示等级关系有助于扩大或缩小检索范围、提高族性检索能力。标题法对于等级关系的揭示所采用的参照符号是“参见和参见自”(See Also and See Also from)，“参见”参照符号后指出的是下级标题词，“参见自”参照符号后指出的是上级标题词，如：

电子计算机
参见 电子模拟计算机
电子数字计算机
电子模拟计算机
参见自 电子计算机
电子数字计算机
参见自 电子计算机

相关关系是指标题词之间除等同关系、等级关系之外语义相关的一种关系,揭示相关关系有助于扩大检索范围,进行相关信息资源的查找。标题法对于相关关系的揭示所采用的参照符号是“参见和参见自”(See Also and See Also from)，如：

图书馆学	情报学
参见 情报学	参见 图书馆学
参见自 情报学	参见自 图书馆学
肿瘤	吸烟
参见 吸烟	参见 肿瘤
参见自 吸烟	参见自 肿瘤

在标题法中,等级关系和相关关系的揭示都用“参见和参见自”作为参照符号,即对这两种关系没有明确区分,在使用中造成了不便。所以从20世纪80年代开始,一些标题表对

其参照系统进行了改进,如广为使用的《美国国会图书馆标题表》自1988年开始改用与叙词法类似的参照符号,明确区分了等级关系和相关关系,使得主题词之间的关系更加清晰。

另外,标题法为了节省标题表的篇幅,标题表一般采用设置复分表的方式,以便对标题词进行细分,其作用与体系分类法的复分表相似。复分表一般包括通用复分表和专用复分表,通用复分表包括地区复分表、时代复分表、文献类型复分表等;专用复分表包括人物标题复分表、疾病标题复分表等。标题表中的复分表实际上也是一种组配方法。所有的复分表一般都只能作为构成子标题使用,唯有地区复分表,还可作为主标题使用。下面是人物标题复分的例子,左边是复分表的一部分细目,右边是标引实例:

— 主义	孙中山—诞辰
— 评论	— 传记
— 诞辰	— 著述— 目录
— 传记	— 学说
— 逝世	
— 遗物	
— 著述	
— 目录	
— 稿本	
— 格言	
— 学说	

综上所述,可将标题法的特点归纳为:①按主题(文献论及或涉及的事物或对象)集中文献;②用经过规范化的语词直接标引文献主题;③用参照系统间接显示主题词之间的相互关系;④用字顺序列直接提供主题检索途径;⑤具有较高的专指度,标题语言的标识先组度高,标识含义比较明确,易于标引,易于检索;⑥有较强的适应能力。标题语言可以随着社会的进步、科学技术的发展自拟新的标题,来表达信息资源中所论及或涉及的新事物,因为标题表中的标题具有示范性,而且自拟或新增标题对标题表的结构没有太大影响。

但是,标题法也存在一些不足,主要表现为:①标题表采用列举方式,往往造成收词量巨大,修订量大的问题。②大量采用定组式标题,使用手工检索工具时只能从规定的组配顺序入手进行查找,无法从多个因素、多个角度进行检索,必然会影响检索效果。③存在集中与分散的矛盾,主要表现在两个方面:一是由于标题法以事物为中心集中文献,从学科或专业角度看,造成了文献资料的严重分散;二是由于标题法用无等级性的词作为标识,标题按字顺排列,所以即使同族事物有时也难于集中。为了弥补文献资料被分散的缺陷,标题法虽然采用倒置标题、多级标题等措施进行弥补,但并不能从根本上解决集中于分散的问题。

作为一种传统的主题法类型,标题法开创了主题法的最初形式,探索了标题法词汇控制的一系列方法,如制定了标题的选择和确定的准则、规定了标题的形式、标题之间关系的揭示、标题标引过程中组配方法的使用等,为主题法的发展奠定了坚实的基础。

(2) 元词法

随着现代信息资源数量的剧增,标题法作为一种列举式主题法类型,已经无法满足信息资源标引和检索的需要。元词法就是为了克服标题法的不足而发展起来的一种主题法类型,它是以元词作为主题标识,通过字面组配的方式表达信息资源主题的主题法。

所谓元词,是指用来标引信息资源主题的、最基本的、字面上不能再分的语词。例如,“化学”、“经济”、“网络”、“图书馆”就属于元词,而“知识组织”、“网络经济”就可分解

为“知识”、“组织”、“网络”、“经济”4个元词。元词的特点是它们在概念上不能再分解，如果再分解便不能表达专业概念，失去检索意义。

元词法是后组式主题法，在使用元词法的情况下，对复合主题资源的标引和检索都是通过元词的组配进行的。例如，“生物信息检索”这一主题，就必须通过“生物”、“信息”、“检索”3个元词的组配进行标引和检索。正是因为元词法是后组式的，所以它只适用于标识单元方式检索系统，如比号卡、比孔卡及书本式检索工具等。因此元词法检索系统必须由两个部分组成：元词卡部分和文献题录卡或文摘卡部分。元词卡部分是在每个元词卡上标引含有该元词的所有文献的文献号，所有元词卡按元词标识的字顺排列，组成检索系统。文献题录卡或文摘卡上著录文献的详细信息，按文献号的顺序组织排列。例如，上述“生物信息检索”这一主题，信息组织时应同时在生物、信息、检索3张元词卡上记录该文献的文献号。同时，当需要检索有关“生物信息检索”的信息资源时，首先应将该检索课题的主题分解为生物、信息、检索3个元词，在检索系统中找到这3个元词卡，然后比对这3个元词卡上的文献号，凡是共同的号码即为符合检索要求的信息资源。

在元词法中，为了避免不同主题的元词之间可能产生的误组配，可使用关联符号加以解决。关联符号是用于揭示同一文献中不同主题概念之间联系强度的一种专用符号，是后组式检索系统中常用的句法手段之一，其方法是在主题词后加上数字或字母符号。元词法还使用职能符号以明确主题标识的关系意义，职能符号是一种表示主题标识在组配中的句法职能的辅助符号，也是后组式检索系统中常用的句法手段之一。如可以规定凡遇到同一篇文献中几个主题词的组配可能产生两种或两种以上含义时，可采用职能符号“A”表示动作对象，符号“C”表示施动者等语义符号，加在主题词后，以防止组配中的误检。

综上所述，可将元词法的特点归纳为：① 标引专指度高，通过多个元词的组配可以表达较专指的概念；② 便于从不同主题角度检索，每个组配的元词都是一个检索入口，可实现多途径检索；③ 检索灵活，利用对元词的增减或改变可以自由地扩大、缩小、改变检索范围。

但是，元词法也存在一些不足，主要表现为：① 直接性差；② 不适宜查找基本主题的信息资源，如不适宜对论述政治、生物等基本主题的信息资源的检索，因为检索到的信息可能是专论政治、生物某一方面的信息的；③ 采用字面组配方式，在字面分解与语义分解不一致时，容易造成误解，如“蘑菇战术”按照元词法的原理应采用“蘑菇”和“战术”两个元词组配，“猎户星座”应采用“猎户”和“星座”两个元词组配，但单独理解“蘑菇”和“猎户”时可能会产生误解，使得信息资源归入不相关的语词之下；④ 早期的元词法不建立参照系统，不显示元词之间的关系，无法进行相关信息资源的查找。

由于元词法存在的这些问题，所以元词法目前已经基本上被叙词法所取代。但元词法在主题法发展中的主要贡献是：率先探索了后组式检索方法，元词法使用的反记法是目前机械检索系统中倒排档的先声，后来为叙词法等主题法类型所采用。此外，元词法还探索了后组式检索中的规律和问题，包括联号、职号等辅助符号的使用方法及对各种检索系统的适应性，为叙词法的发展和使用开辟了道路。

（3）叙词法

叙词法是以从自然语言中精选出来的、经过严格控制的语词为信息资源的主题标识，通过概念组配方式表达信息资源主题的主题法类型。叙词，国内亦称主题词，是经过规范化处理的、以基本概念为基础的表达信息资源主题的词或词组。叙词语言是受控主题语言的主流，到目前为止，国外的叙词表不少于千种，国内也超过130种，我国目前使用最广泛的叙词表是《汉语主题词表》。

叙词法集多种检索语言的优点于一身,是多种检索语言的原理和方法的综合,使其成为一种具备优异性能的现代检索语言。例如,词的组配,来自单元词法;概念组配,来自组配分类法;适当采用预先组配(词组),来自标题法;对语词进行严格的规范化处理,来自标题法、单元词法;用参照系统显示词间关系,来自标题法;编制范畴索引和词族索引,来自体系分类法;编制叙词轮排索引,来自关键词法;以语词作为标识和按字顺排列,来自标题法和元词法。

在叙词法的基本原理中,概念组配是叙词法最基本的原理,叙词法以概念组配代替元词法的字面组配,两者存在以下不同。

第一,语词单元不同。元词法严格采用字面上不能再分的语词为标识单元,有时会影响对其主题内容的确切表达,产生误检,而概念组配要求词汇单位既能组配表达概念,又能独立表达概念,而且在这两种情况下表达的概念应该一致,因此用于概念组配的词汇可以是元词,也可以是词组,对主题的表达更准确。如字面组配时,“隧道二极管”可以分拆为“隧道”和“二极管”,然后用于组配,但是在概念组配中,由于“隧道”作为组配因素所表达的概念和它独立表达的概念不一致而会造成误检,因此,需直接使用“隧道二极管”做叙词,无须分解。再如,在标引“美术学校”这一主题时,元词法使用“美术”、“学校”两词组配,会出现美术专科学校和学校的美术课两种含义,而叙词法直接以美术学校进行标引,就不会出现二义性。

第二,组配的原则方法不同。概念组配本质上是在概念分析的基础上进行概念综合,即概念组配结果所表达的概念与参与组配的各方所表达的概念应符合概念逻辑原理,一般表现为下位概念与上位概念的关系,字面组配则利用构词法进行分拆和组合,它符合构词规律,但是不一定符合概念逻辑。如“生物物理学”这一主题,按照字面组配,可采用“生物”、“物理学”加以组配。但根据概念组配原理,则应使用该主题构成的概念单元,以“生物学”和“物理学”进行组配,显然后者比前者的表达更准确。所以字面组配和概念组配有时在字面形式上一致,有时不一致。一致仅仅是因为字面组配的构词结果和概念组配结果的字面表达正好吻合,并不说明两者在本质上一致。

同标题法一样,叙词法也通过参照系统揭示词间关系,即通过在叙词下设置参照项的方法,在叙词之间建立起一种反映主题词之间联系的语义网络。叙词之间的关系主要有三种:等同关系、等级关系和相关关系。

等同关系是指正式叙词与非正式叙词之间的关系,亦即在词汇控制中被选作叙词的词与落选且保留在叙词表中作为非正式叙词的词之间的关系。在叙词法中显示等同关系使用的参照符号为:用(Y或USE)和代(D或UF)。参照符号“Y”后指出的是正式叙词,“D”后指出的是非正式叙词,正式叙词可用于信息资源的标引和检索,非正式叙词不能用于信息资源的标引和检索,只起检索入口作用,但需保留在叙词表中。例如:

原子能工业	核工业
D 核工业	Y 原子能工业
毕业实践	毕业设计
D 毕业设计	Y 毕业实践

等同关系除用于揭示同义词、准同义词之间的关系外,还用于揭示组代关系,即指示一专指的非叙词和相应的叙词组配形式之间的关系。例如:

国际货币制度危机
Y 国际货币制度 + 货币危机

等级关系是指上位概念主题词与下位概念主题词之间的一种关系,对等级关系的揭示有

助于扩大或缩小查找范围。在叙词法中显示等级关系使用的参照符号为:

分项	F	NT
属项	S	BT
族项	Z	TT

参照符号“F”后指出的是下位概念主题词,“S”后指出的是上位概念主题词 “Z”后指出的是族首词,族首词是一族词中最泛指的上位词。例如:

农业政策

F 副业政策

粮食政策

渔业政策

S 经济政策

Z 政策

相关关系又称类缘关系,是叙词之间除了等同关系、等级关系之外语义相关的一种关系。相关关系是用于揭示叙词之间的各种联系、扩大检索范围、进行相关资料查找的主要手段。

在叙词法中显示相关关系使用的参照符号为: 参项 C RT

例如: 教育思想

必要劳动

C 教育理论

C 剩余劳动

对外贸易

分类表

C 国际贸易

C 分类号

综上所述,叙词法的特点可归纳为:①词汇控制严格,可以根据检索系统的需要对词汇进行有效控制;②组配准确,标引能力强,能够准确、专指地标引和揭示各种主题内容;③检索效率高,可以通过灵活组配方式进行多途径检索;④对检索系统适应能力强,可以同时适用于标识单元检索方式和文献单元检索方式,既能较好地适应计算机检索系统的要求,又能适应于手工检索系统的需要。

叙词语言的不足:①由于词汇控制要求严格,词表编制和管理的难度大;②文献标引须在概念分析的基础上进行,标引难度大,要求高。

(4) 分类主题一体化语言。

分类法与主题法的有机结合就是分类主题一体化,分类主题一体化实际上有两层含义:其一是指检索系统或检索工具的分类主题一体化;其二是指检索语言的分类主题一体化。本书主要指后者。分类主题一体化语言是一种实现了分类语言和主题语言兼容互换的系统。即一个分类系统与一个主题系统实现了完全兼容,有机地融合为一个整体,既能充分发挥各自独特的功能,又能通过配合发挥最佳的整体效应。

分类法和主题法的发展说明了二者在相互渗透和相互融合,如分类表中编制字顺索引,主题词表中编制分类索引、词族索引等,但它们仍不能实现两种检索语言的有机结合,即分类主题一体化,它们都只具有单一的分类标引或主题标引的功能。20世纪60年代中期以后,国外对分类表和主题词表进行了大量的抽样调查试验,并在此基础上开始了分类法主题一体化的研究,这些调查分析表明,分类法和主题法之间有着非常密切的对应关系,在此基础上可以实现分类语言与主题语言之间的兼容和互换。分类法和主题法虽然有着种种差异,但是它们在原理上却有着许多相同之处,正是这些相同之处成为它们结合的基础。而且随着电子计算机在图书馆和文献信息工作中的广泛应用和情报语言学研究的不断发展,使分类主题一体化词表的产生成为可能。

分类主题一体化语言具体体现为分类主题一体化词表的编制,一体化词表的主要组成部

分是分类表部分和主题词表部分,它们是一个统一系统中相互配合、又相对独立的两个子系统。这两个子系统通过术语的统一、词间关系显示的互补及同一的标记作为连接件或转换器而有机地结合为一个整体。在一体化词表中,分类表部分的功能优于单一的分类表,主题词表部分的功能也优于单一的主题词表,而且它们通过内部协调,分类表和主题词表各自提高了自己的特性和功能,使得一体化词表的整体功能高于它的各个部分(分类表、主题词表)功能的总和。

分类主题一体化词表的类型主要包括:① 分面叙词表,这是最典型的分类主题一体化词表,是由一部分面或半分面的分类表和一部叙词表组成的,有的还附有轮排索引及英汉对照索引,如《教育主题词表》、《社会科学检索词表》等;② 分类主题词表,又称分类法—主题词表双向对照索引,如《中国分类主题词表》等;③ 集成词表,它是将特定领域的若干叙词表和分类表汇编而成的一种集成词表,用于联合分类标引和主题标引,实现分类语言与主题语言之间的兼容及互换,如中国医学科学院编制的《中图法R类与MeSH、中医药学主题词表对照表》就是一部以分类表(《中图法》的医学类)为主干,与两部主题词表(国际医学界通用的《医学主题词表》和我国的《中医药学主题词表》)进行对应标引而形成的集成词表。

2) 按语词标识的组配特点划分的类型

(1) 先组定组式主题法。

是指复杂主题的标识,在词表中就已经组配好了,使用时可以直接从词表的标识中选取。标题法就属于这一类型。

(2) 先组散组式主题法。

是指复杂主题的标识,在词表中未组配好,而是在标引阶段根据信息资源的主题需要进行组配的。采用叙词表在标引阶段建立标题,就属于这一类型。

(3) 后组式主题法。

用户检索前,主题系统中的主题词是单立的,检索时才根据检索需要进行组配。如在检索“湖泊水污染”这一主题前,检索系统中只有“湖泊”、“水污染”等单立的主题词,用户输入检索要求后,检索系统经过匹配,才形成“湖泊—水污染”的组配标识。元词法、叙词法就属于这一类型。

3) 按是否对主题词进行控制划分的类型

(1) 受控主题法。

受控主题法指依据特定主题词表揭示信息资源的组织方法。如标题法、叙词法等均属于这一类型。它们的共同特点是标引和检索均依据选定的主题词表对主题概念进行转换,从而可以通过词表对信息资源内容的规范表达和相互关系的揭示来改进检索效果。

(2) 非受控主题法。

非受控主题法即自然语言检索系统,是直接使用信息资源或用户检索使用的自然语言语词进行组织的方法,包括关键词法、自然语言文本检索等。这种方法的特点是不需要使用主题词表,但一般仍需要遵守一定的标引规则和检索措施,以改进使用效果。

3.3.2 国内外常用主题词表介绍

1. 《美国国会图书馆标题表》

1) 概况

《美国国会图书馆标题表》(LCSH, Library of Congress Subject Headings)是世界上最具影响的一部标题表,是美国国会图书馆在编目实践的基础上编制而成的。该表于1909—1914

年以《美国国会图书馆字典目录用标题表》名称陆续出版了7版,1975年的第8版改为现名,并在同年出版缩微版,1986年以机读形式出现,称为主题规范档,同年出版第10版,自1988年第11版起,每年出一新版。2001年出版网络版(Classification Web)(<http://classification.web.net>)。该表是目前美国及全世界使用最广泛的标题表,美国国会图书馆发行的印刷卡片、机读目录及西文图书的在版编目数据都依据LCSH的标题进行编制,而且世界上其他许多国家图书馆在对英文图书编目时,也基本以该表或参考其编目数据进行主题标引。

LCSH 现有三种基本形式:印刷版、机读版、缩微平片版。它的管理工作由美国国会图书馆的主题编目部完成,每周更新,平均每年大约新增5000~7000个标题。为了及时规范和反映标题的动态,还有如下4种辅助工具。

(1)《主题编目手册:标题表》(手册),1984年首次出版,现用1991年的第4版,该手册对标题表的建立和使用作了详细说明。

(2)机读版的“名称规范档”,LCSH不列举大多数专有名词,要求直接使用名称规范档的相应名称作为标题。

(3)《编目服务通报》,收录《手册》出版后标题的修改、更新、使用及有关出版物的信息。

(4)《自由浮动复分标题字顺索引》,将《手册》中按范畴编排的全部自由浮动复分标题按字顺编排,提供字顺查找途径。

LCSH 由主表、副表和使用说明三部分组成。其中,主表是标题表的主体。目前其印刷版分为5卷,包括导言和字顺表,其副表和使用说明收入《主题编目手册:标题表》中。

2) 款目结构

LCSH 的主表是由众多标题款目和非标题款目按字顺排列而成的。

(1) 标题款目的结构。

标题款目包括以下几个部分。

① 主标题,用黑体印刷,作为款目词决定该款目在标题表中的位置,是标题表进行主题标引和检索的依据。标题形式有单词标题、词组标题、倒置标题、多级标题、带限义词的标题等。

② 分类号,置于方括号内,放在主标题之下,这是与该标题对应的美国国会图书馆分类法的分类号。不是所有的标题下都有分类号。如果一个标题可分入一个以上的类,它会有多个分类号,有些分类号后还注明其学科属性,但该分类号不能用于分类标引。

③ 注释,放在标题或分类号之后,许多标题有范围注释,用以说明该标题的适用范围或几种含义及与其他相关标题的界限。此外,有些标题后的“May Subd Geog”或“Not Subd Geog”可以算是地理复分注释。

④ 参照项,用于揭示标题之间的等同关系、等级关系和相关关系。1986年前使用的参照符号有:See、X、SA、XX。揭示等同关系的参照符号是See和X,揭示等级关系和相关关系的参照符号都是SA和XX。从1986年12月的LCSH平片版和1988年的第11版开始,LCSH改用与叙词法相似的参照项目及符号,并明确区分等级关系和相关关系,使词间关系更清楚,使用的参照符号有:UF、BT、RT、NT,同时保留原来使用的SA(See also)作说明参照,用于揭示一组相关标题或指示查找相关标题的方法。

⑤ 复分标题,也用黑体印刷,前面用“—”表示。复分标题下可以有复分标题,前面用两个短横表示,复分标题之下也可能有参照项。标题款目样例:

Agricultural machinery (May Subd Geog)

[S671-S760.5]

- UF Agricultural —Equipment and supplies
 - Crops—Machinery
 - Farm machinery
- BT Machinery
- RT Farm—equipment
 - Farm modernization
 - Machine—Tractor stations
- SA subdivision machinery under names of crops ,
 - e.g. Corn—Machinery
- NT Agriculture engineering
 - Agricultural instruments
 -
- Cost of operation
 - UF Agricultural—Operating cost
- Dynamics
 - BT Dynamics
- Electric equipment
 - BT Electricity in agriculture

(2) 非标题款目的结构。

非标题款目包括以下几个部分。

① 非正式标题，不用黑体字，依据其字顺与正式标题混排。

② USE 参照项，这是必有的一项，它可以指向一个、几个或一批正式标题。

非正式标题款目样例：

Cars (Automobiles)
USE Automobiles

3) LCSH 中复分的使用

西文书刊一般以 LCSH 作为主题标引的工具。主题标引除了将直接从 LCSH 中查到的单词标题或短语标题等作为主标题外，多数主题还要通过复分进行组配标引。复分有以下几种类型。

(1) 论题复分，用在主标题或其他复分下，将标题所表达的概念限定在专门的子论题或子方面，在 USMARC 中记录在 6XX 主题字段的 \$x 子字段，如 650#0 \$a Computer networks \$x Security measures。

(2) 地理复分，用以指明主要论题的起源或地理位置，在 LCSH 的词表中凡标注 (May Subd Geog) 的标题均可进行地理复分，若在某个主标题及复分标题后均出现 (May Subd Geog) 时，地理复分应在相应的复分标题后出现。地理复分在 USMARC 中记录在 6XX 主题字段的 \$z 子字段，如 650#0 \$a Cameo glass \$z Ohio \$z East Liverpool。

(3) 年代复分，用以限定主标题或复分标题所反映文献内容涉及的历史时代。年代复分在 USMARC 中记录在 6XX 主题字段的 \$y 子字段。年代复分一般不能自由浮动，而要严格按照主表中在不同标目下所设置的年代复分来标引，如 650#0 \$a Costume \$z France \$x history \$y 17th century。

(4) 形式复分，用以指明文献资料主题的形式特征，形式复分在 USMARC 中记录在 6XX 主题字段的 \$v 子字段，一般位于主题标引的最后一级。常见的形式复分有：Abstracts；

Bibliography; Catalogs; Congresses; Dictionaries; Handbooks; Periodicals 等。例如: 650#0 \$a Calvinism \$z Switzerland \$v Bibliography。

(5) 衍生复分, 在 LCSH 中的表示方法是: 将复分中的某一词用方括号括起, 指明该词在针对具体文献标引时可用其他同类词置换并取消括号。例如: World War, 1939-1945--Personal narratives, American[French, German, ect.]

根据这个衍生复分模式, 可以派生出以下标题:

650 #0 \$a World War, 1939-1945 \$x Personal narratives, Spanish

(6) 自由浮动复分, 为了便于在主题标引时对标题进行有效控制, LCSH 使用自由浮动复分的形式进行组配。自由浮动复分标题大多有一定的应用范围, 在 LCSH 的手册 (Subject Cataloging Manual: Subject Headings) 中有具体说明, 一定要按照规则使用。自由浮动复分包括单独列表和采用类型标题的形式。单独列表又包括通用自由浮动复分表和特定范围自由浮动复分表。

通用自由浮动复分表由列举的形式子标题和论旨子标题构成, 用于相关类目的复分, 此表收入《手则》中, 使用范围一般在各子标题下注明。例如:

高度 (Altitudes)

用于地区、国家、州等的地名下。

见自 海拔

历史 (History)

用于各类主题, 包括地区、国家、城市 等名称之下。

特定范围自由浮动复分表收入只适用于特定范畴的标题, 包括人物、种族、团体、人名、家族名、地名、水体等, 集中在《手则》中, 供相应主题选用。例如, 用于地名下的复分标题有:

Abstracting and indexing

Abstracts

Air defenses

.....

类型示范标题表, 即为各类标题分别选定一两个示范标题, 并在其下列出一整套标准化、规范化的适用于这类标题的复分词, 供同类标题仿照使用, 这种复分有如分类表中的仿分。

LCSH 在 35 个类下共设置 43 个示范标题。例如:

Catalog (类别)

Animals (General) (动物)

Animals, Domestic (家畜)

Chemicals (化学制品)

Colonies (殖民地)

Educational institutions (教育机构)

Individual (个体)

Types (类型)

Industries (行业)

.....

Pattern Heading (示范标题)

Fishes (鱼)

Cattle (牛)

Copper (铜)

Insulin (胰岛素)

Great British-Colonies (英国一殖民地)

Harvard University (哈佛大学)

Universities and Colleges (大学与学院)

Construction industry (建筑业)

Retail trade (零售业)

.....

(7) 复分的顺序。在组配涉及多种主题成分时, 主题标目的各种复分顺序一般为: 主标

题—论旨子标题—地区子标题—编年子标题—形式子标题,但有时地区子标题也可置于论旨子标题前,通常由标题词后的地区复分指示规定。即凡是主表中在主标题词后注明采用间接法,复分的次序为:主标题—地区子标题—论旨子标题—编年子标题—形式子标题。如主表中在副标题后注明采用间接法,复分的次序则为:主标题—论旨子标题—地区子标题—编年子标题—形式子标题。例如:

Construction industry (May Subd Geog)

—Finance

—Law and legislation (May Subd Geog)

—Government policy (May Subd Geog)

—Industry capacity

如用 Italy 复分,则可产生下列地名复分标题:

Construction industry—Italy

Construction industry—Italy—Finance

Construction industry—Finance—Law and legislation—Italy

Construction industry—Government policy—Italy

Construction industry—Italy—Industry capacity

4) LCSH 简评

LCSH 能在世界上许多国家广泛流行,是因为它具有以下的特点:① 美国国会图书馆在其发行的印刷卡片、机读目录和西文图书的在版编目数据上都标有 LCSH 的标题,扩大了影响,使其成为国外使用最广的主题词表;② 它较好地体现了克特关于标题法的理论,确立了主题法的一些基本原则,如标题的选词原则、标题词的形式,明确规定标题的复分方法,建立参照系统等,便于利用;③ 它是以美国国会图书馆藏书的实际需要为基础而编制的,学科面广,标题详细,所建立的标题有充分的文献保障,有很强的实用性;④ 它有专门的管理机构,由美国国会图书馆编目部负责定期修订,使其具有长久的生命力,并备有《手册》等配套工具书,保证其使用的一致和规范。

但是 LCSH 也有一些不足,主要表现在:① 缺乏统一的理论指导,不同时期在标题方式和形式(如同类标题在使用正写与倒置、短语与加副标题、单数与复数等方面)的处理中存在不一致;② 参照不严密,缺乏规律性和一致性,许多款目未做互逆参照,不少款目在从传统标题参照形式向叙词参照形式转换时未准确揭示关系类型;③ 缺乏专指度,采用先组方式无法充分标引较专指的信息资源;④ 社会科学领域的标题带有强烈的美国中心和政治、社会等方面意识形态的影响,对于有不良倾向的某些主题须做修改。例如,China—history—Taiping rebellion, 1850—1864 应改为 China—history—Taiping Uprising, 1850—1864; Li Tzu Ch'eng Rebellion, 1628—1645 应改为 Li Zicheng Uprising, 1628—1645。

2. 《汉语主题词表》

1) 概况

《汉语主题词表》(以下简称《汉表》)是“汉字信息处理工程”的配套项目,1975 年开始编制,由原中国科技情报研究所和原北京图书馆主持编制,1980 年出版。它是一部大型的综合性主题词表,全书共分 3 卷 10 册,第 1 卷是社会科学部分,包括 2 个分册(第 1 分册为主表,第 2 分册为索引);第 2 卷是自然科学部分,有 7 个分册(第 1~4 分册为主表,第 5 分册是词族索引,第 6 分册是范畴索引,第 7 分册是英汉对照索引);第 3 卷是附表,只有 1 个分册。整个词表共收主题词 108 568 条,其中正式主题词 91 158 条,非正式主题词 17 410 条。

为了使《汉表》跟上时代发展的要求,1991年,原中国科学技术情报研究所对词表的自然科学部分进行了修订,出版了自然科学的增订版,并建立了词表管理系统。1996年,根据增订版补充编制了自然科学部分的轮排索引,从而使词表结构更加完备。目前《汉表》有印刷版、机读磁带版两种形式。自1985年起,《汉表》的主题词已被北京图书馆用于统编卡片的主题标引,我国的综合性图书馆现在一般均以《汉表》作为信息资源主题标引的工具。

2) 结构

《汉语主题词表》由主表、辅助索引、附表组成。主表是词表的主体,作为信息资源标引和检索的依据;附表主要收录专有名词;辅助索引是通过改变组织方式,提供从不同途径着手查找叙词的工具,包括范畴索引、词族索引、轮排索引和英汉对照索引4种。

(1) 主表。

主表又叫做字顺表,是由众多叙词和非叙词按字顺排列而成的,是信息资源标引、检索及组织文献数据库和目录索引的主要工具。《汉表》作为一部大型的综合性叙词表,收词范围包括社会科学、自然科学各学科领域的主要名词术语。按照词形结构,主表的叙词可分为后组词和先组词两种形式。为了使主题标引能够同时适合机械和手工标引检索的需要,词表在大量选用先组词的同时,适当放宽词组的选择范围和级别,增加了词组的数量。选词主要遵循以下原则:选定的主题词应是各学科领域信息资源中经常出现的、在信息检索中有使用价值和一定的使用频率、能作为主题汇集一定量信息资源或具有叙词组配功能的名词术语;选定的主题词必须词形简练、词义明确,严格遵守一词一义原则,并且通过概念组配能表达信息资源或用户查询的特定主题;选定的主题词应符合我国科学发展的实际需要,尽量与国内外主要主题词表相兼容,并应注意主题词的科学性与思想性。

正式主题词(叙词)的款目结构由汉语拼音、款目主题词、范畴号、含义注释、英文译名和各种参照项组成,正式主题词款目内的主题词一律用黑体排印。非正式主题词款目内的非正式主题词一律用白体排印,但其参照项内所指引的正式主题词则一律用黑体排印。正式主题词款目样例:

汉语拼音	————→	Xian Xiang Guan	
款目主题词	————→	显像管	[56E] ←—— 范畴号
黑体)			
英文译名	————→	Picture Tubes	
代项符号	————→	D 电视显像管	←—— 非正式主题词
		监视管	
分项符号	————→	F 彩色显像管	
		固体显像管	←—— 下位主题词
		黑白显像管	
属项符号	————→	S 电子束管	←—— 上位主题词
族项符号	————→	Z 电子管*	←—— 族首词
参项符号	————→	C 显示管	←—— 相关词
		指示管	

非正式主题词(非叙词)款目样例:

汉语拼音	————→	Nongye Fangzhen Zhengce	
非正式主题词	————→	农业方针政策	[05A] ←—— 范畴号(白体)
英文译名	————→	Agricultural Program and Policy	
用项符号	————→	Y 农业政策	←—— 正式主题词(黑体)

(2) 附表。

附表是从主表中析出的几种专用词汇表。它所收录的主题词包括世界各国政区、自然地理区划、组织机构和人物等4个范畴领域中比较重要的专有名词。这些主题词一般都具有单独概念的性质,以及较强的检索意义或组配作用。《汉表》的附表包括:

附表一,世界各国政区名称(1100词),收入世界各国、地区及所属重要城市名称;

附表二,自然地理区划名称(361词),收入世界重要自然地理区划名称,包括山、川、河、流、湖、海、洋、岛屿、平原、盆地等的名称;

附表三,组织机构名称(1900词),收入各专业中具有研究价值和文献论述的重要机构团体名称,但关于政治派别、军队中外历史上的机构团体名称均收入主表。

附表四,人物(4765词),古今中外具有研究价值和文献论述的人物对象。

附表主题词的款目结构和排列方式与主表基本相同,其款目均由汉语拼音、汉语名称和英译名构成,有的款目下还有注释和参照项,分别按汉语拼音字母顺序排列。这类主题词一般不收入范畴索引,因此其款目项中没有范畴号。

(3) 辅助索引。

《汉表》的辅助索引是通过改变组织方式,提供从不同途径着手查找叙词的工具,包括范畴索引、词族索引、轮排索引和英汉对照索引4种。

① 范畴索引,是将主表的全部叙词按词义所属范畴划分成类,按类目关系编排的词汇分类体系。它共设58个大类,675个二级类,1080个三级类,其中社会科学15个大类,173个二级类,311个三级类;自然科学43个大类,502个二级类,769个三级类。用两位数字和两位字母作为类目标记。范畴索引允许某些词同时归入2个范畴甚至3个范畴。索引中,非叙词下一律以“Y”项指引与之对应的叙词。族首词后一律缀以“*”号,作为联系查找词族索引的符号。范畴索引样例:

07	文化事业	←	一级类目
07A	文化事业一般概念		
07B	社会文化工作	←	二级类目
07C	出版、发行		
07CA	出版、发行一般概念		
07CB	出版编辑	←	三级类目
07CC	出版物类型		

② 词族索引,是把主表中具有属分关系、某些整体与部分关系的正式主题词,按规定属分级别展开全显示的一种词族系统。这种索引在机检系统中是实现自动扩检、上位词登录及满足族性检索的重要手段,同时又是在标引和检索中提供系统性查词和选定标引词的辅助工具。族首词是指能概括一族主题词的最上位的广义概念词,它列在一族之首,词后附有“*”号以示区别。索引中以族首词作为排检款目词,按汉语拼音音序排列,族首词下的主题词按其概念广狭的等级阶梯式排列,其等级关系用“.”的数目表示。族首词为一级词,其下以前置一个点“.”为二级词,两个点的“..”为三级词,三个点的“...”为四级词,依此类推。全表共收词族3707个,其中社会科学886个,自然科学2821个。自然科学增订本新增词族141个,删除族首词134个,将词族调整为2825个。最大词族收词达5270个,最少词族收词只有2个。词族索引样例:

```
shi ge
诗歌
. 讽刺诗
```

- . 格律诗
- . 古典诗歌
- . 古体诗
- .. 七言古诗
- .. 四言诗
- .. 五言古诗
- .. 杂言诗
- . 近体诗
- .. 绝句
- ... 七言绝句

③ 轮排索引, 目前《汉表》的轮排索引只包括自然科学部分, 收入《汉表》自然科学部分的全部词汇, 包括正式主题词和非正式主题词。采用题内关键词索引的方式编制, 检索入口位于索引款目的中部, 保留位于检索入口左侧及右侧的上文及下文, 索引地址为范畴号和页码位置号。正式主题词索引款目附设代项 (D), 非正式主题词款目附设用项 (Y)。当正式主题词的词首词素作为检索入口时, 加注了该词的英文。为了便于区分, 正式主题词用黑体印刷, 非正式主题词用白体印刷。每个正式主题词或非正式主题词根据其构成词素, 即有检索价值的单元概念, 确定轮排数量。条目的排序采用双向排列法, 即先按检索入口右侧汉字的拼音排列, 右侧相同的主题词, 则按检索入口左侧的汉字从右向左排, 以便将靠近入口词的汉字有层次地排列, 方便用户选择使用。

轮排索引是将主表中全部正式主题词和非正式主题词按其词素予以轮排, 将含有同一词素的正式主题词和非正式主题词聚集在一起, 如果用户记不准某一主题词的确切词形, 也可以从某一词素出发查得该词, 并同时找到包含该词素的全部正式主题词及非正式主题词。例如, 靶场安全、工厂安全、导弹安全、反应堆安全、激光安全、核安全、安全棒、反应堆安全保险装置等词, 在主表中被分散在不同的字顺位置, 但在轮排索引中, 可通过相同的词素“安全”聚集在一起。只要按排检位置就可找到“安全”这一检索入口, 即可查到包含这一词素的所有正式主题词和非正式主题词。轮排索引样例:

	检索入口	范畴号及其用代参照	主表页码及栏目
	安全*	83NA	14 右
	Safety		
靶场	安全	83NA	26 中
工厂	安全	83NA	912 右
行车	安全	83NA	2823 中
导弹	安全	83NA	417 右
反应堆	安全	52DG	689 左
激光	安全	56DA 83NA	1244 左
核	安全	52HB Y 辐射防护	1069 右

	安全棒	52DE D 紧急停堆棒, 事故棒	14 右
	Safety rods		
反应堆	安全保险装置	52DE	689 中

④ 英汉对照索引, 这是汉语主题词表的辅助工具, 供标引人员和检索人员从英文反查

汉语主题词之用。一般将每个主题词包括正式主题词和非正式主题词都尽量译成英文,非正式主题词后以“Y”项列出相应的正式主题词。原则上,一个汉语主题词仅选择一条与之对应的英文译名。英汉对照索引样例:

Carcinoma
癌
Carcinoma in Situ
原位癌

3)《汉语主题词表》评价

《汉表》是目前我国规模最大的一部叙词表,是我国情报检索语言发展史上的一个重要里程碑。其特点可归纳为以下几点。

① 结构完备。由主表、附表、索引多个部分组成,词汇控制严格,整体功能完备,是传统词表编制的一种典型模式。

② 词汇丰富。《汉表》的词汇丰富,选词遵循思想性、科学性、实用性、兼容性的原则,参考了当时国内外多种词表,并结合信息资源标引的需要,涉及范围广,专指度深,首次为我国文献信息单位提供了一个权威的词汇集合。

③ 探索了词表编制方法。通过汉语主题词表的编制,不仅总结了国内外词表编制和词汇控制的技术方法,而且还结合实践进行了广泛探索,为汉语叙词表的发展积累了经验。

④ 探索了叙词表的实际使用方法。结合《汉表》的特点和实际使用的需要,我国标引工作者探索了在手工检索系统中 and 计算机检索系统中的使用形式,推动了叙词法在我国的实际使用。

⑤ 国家图书馆在发行的印刷卡片、机读目录及再版编目数据上都使用《汉表》的主题词,是国内文献单位通用的综合性主题标引工具。

但是《汉表》也存在一些不足,主要表现在:

① 结构上,印刷型词表因篇幅巨大,组成分散,编排不紧凑,整体性较差。

② 词汇上,存在个别专业收词过多,不同专业之间收词数量不平衡的问题,一些常用术语和反映学科动态的词没有收入,造成《汉表》词多而不精,面广而不匀。而且《汉表》先组词偏多,词组占 60%以上,元词所占比例较少。

③ 在词间关系处理上,等同率(非叙词与叙词的比率,等同率越高,提供的非叙词多,检索入口更多,使用越方便)低于国内外同期专业词表的平均值,基本上不设与组代词的等同关系项,无关联词数量大,使词表易用性较差。

④ 没有设置统一的管理机构,未确定修订方针,缺乏健全的管理机制,自 1979 年词表出版以来,仅在 1991 年对自然科学部分修订过一次,社会科学部分只在《中国分类主题词表》编制时对词汇进行了必要增补,由于修订周期过长,无法满足实际需要。

⑤ 目前仅自然科学的词表实现了机读化,整个词表还没有完整机读版本,这是与《汉表》的地位和使用的需要不相称的。

《汉表》有如下几个发展方向:

① 《汉表》的国家标准化与专业词表的研制同时进行,处理好专业主题词表与《汉表》的兼容与统一问题,建立起一个主题词表统一和兼容的整体系统。

② 对《汉表》进行改造,如在词汇方面,缩减词汇,减少先组词的比例;在结构方面,简化词表宏观结构,细化词表微观结构,合并词族索引于主表中,词族展开的形式与自动标引中广泛采用的最长匹配法的工作形式相吻合,有利于提高系统效率;在兼容方面,英汉对照索引中的英文词逐渐过渡为英文叙词,并在此基础上发展多语种对照索引;在协调性方面,

保证社会科学部分与自然科学部分的同步发展。通过上述改造使其成为中文自动标引中通用的综合性切词词典。

③ 利用《汉表》的编制成果来构建领域本体,实现叙词表到本体的转换。叙词表和本体都在基于知识理解的基础上构建,都涉及知识的分类及语义关系的构建,二者有融合的前提。但二者构建目的不同,存在一定差异,叙词表作为规范化的术语词表,是为提高计算机的检索效率而制定的,本体以概念和概念之间关系的建立为核心,注重计算机的形式化描述,以计算机能够理解的语义内容为前提。若使叙词表较准确地转换为本体,需要考虑其特点,在进一步的数据整理及语义关系调整的基础上进行。这是一项非常有意义和价值的研究,如果将人工研究的转换规则转换成计算机处理形式,将会大大提高转换效率。未来的研究更应重视采用机器学习的方式构建人工智能系统,实现《汉表》到领域本体的自动转换。

3. 《社会科学检索词表》

1) 概况

《社会科学检索词表》是中国社会科学文献信息中心编制的一部供社会科学文献资源标引和检索的词表,为分面叙词表,是一部分类主题一体化词表。词表的编制自1987年立项开始,到1993年12月结束,前后历时7年,期间经过试标引和反复修改、充实,全表共收词23 000个,其中正式叙词18 874个,非叙词4126个,由分类表、字顺表和英汉对照表三部分组成,是国内供社会科学领域文献单位进行标引和检索的多学科性的工具。

2) 分类表

分类表部分是一部将传统分类法设置特点和分面分类方法相结合而编制的分面分类系统。即将整个社会科学领域区分为17个基本大类,27个专业类目,列举如下:

A 马克思列宁主义	L 法学
B 哲学	M 军事学
BM 逻辑学	N 新闻学、传播学
BN 理学	P 图书馆学、情报学、档案学
BP 美学	Q 教育学
C 宗教学	Y 管理科学
D 语言学	YB 统计学
E 文学	YL 科学学
F 历史学	YM 未来学
G 考古学	YP 管理学
H 社会学	YQ 人才学
HP 人口学	YR 决策学
I 民族学	a/y 通用概念
K 政治学	

在基本大类之下,根据学科特点和需要进一步设置若干基本类,然后采用分面分析的方法设置类目。根据社会科学各门类的情况,分类表将各学科下可能出现的基本分面确定为9个,依次为:历史面、学派面、方法论面、学科面、理论面、结构面(事业、结构、组织机制等)、管理面、时间面、空间面等。每个基本类下的分面不一定完全相同。例如,社会学大类下分为6个面:

H 社会学	
H02 社会学史	历史面
H03 社会学学派与理论	学派面

H05	社会学研究方法	方法论面
H08	社会学学科	学科面
H12	社会心理学	
H2	社会	
H3	婚姻与家庭	理论面
H4	社会管理	
H5	社会服务	管理面
H6	社会问题	

在基本分面之下，再进一步根据分面的处理对象，按照一定的标准进行区分，列出具体的类目。当一个分面下按几个标准展开时，一般通过分面标头的形式揭示其采用的分类标准，例如：

社会角色
按职权分)
社会角色职责
社会角色期待
社会角色权利

.....

(按行为分)

社会角色冲突
社会角色变化

.....

这里，(按职权分)和(按行为分)为分面标头，放在括号里。

下面是一个三级类目分面样例：

H · 社会学
H08 · 社会学学科
 (按研究方法分)
H081.2 ... 应用社会学
H081.4 ... 理论社会学
 (按研究领域分)
H081.6 ... 社会传播学
H081.8 ... 教育社会学
 (按研究范畴分)

.....

其中圆点的数目表示分类等级，一个圆点为一级类目，两个圆点为二级类目，三个圆点为三级类目……依此类推。分类标记采用汉语拼音字母和阿拉伯数字组成的混合号码，基本按照层累标记方式配置号码，类目的等级关系主要应根据类名前以小黑点表示的类目等级。

分类表的编制。首先把分类表中的每个类名叙词化(即主题词化)，也就是对分类表的每一个类名实行严格的词形、词义和词间关系控制，要求一个类名代表一个主题概念，一个主题词(即叙词)只用一个类名表示。例如，“文学”在传统的分类法中只是一个基本大类，但在本词表中，它既是分类表部基本大类的一级类名，又是字顺表的一个叙词。然后按照叙词的分面要求将“文学”分为“文学史”、“文艺理论”、“文艺事业”、“文学翻译”、“文艺批评”、“文学创作”、“文艺作品”等10个组面，作为二级类名归入分类表的基本大类之下。

这 10 个二级类名在字顺表中同时又是 10 个叙词……依此类推, 有的类表最多可达到 8 级类名。这样, 同一个主题概念同时分布在本词表的分类表和字顺表两个部分中, 并用同一个符号加以标识。通过软件控制, 一次检索即可自动生成两种检索标识, 并收到两种检索语言兼容互换的效果。

对分面类表中采用的类名按照叙词的要求进行词汇控制, 即分类表在类名下直接使用主题法的参照符号。

例 1:

E021 ... 世界文学史
解释项 ———> J 世界断代文学史用组配标引, 如:
世界近代文学史 ———> 世界文学+近代

例 2: N165 ... 广告心理学

属项符号 ———> S 传播心理学 N108.1

例 3: PQ65 ... 文献复制

代项符号 ———> D 复制服务

属项符号 ———> S 情报服务 PR7

分项符号 ———> F 档案复制 PS367

参项符号 ———> C 文献复制服务 PQ446

此外, “d/y 通用概念表”收入所有社会科学各类使用的通用概念, 按文献、时间、地域、人员、团体与机构、会议、一般概念的次序排列, 可以根据需要与各门类的主题类目进行组配标引。

通过上述方式, 分类表将类目按照学科与分面的结合组织成一个系统, 用户可以通过分类表进行查找和检索, 同时也可以作为主题词表的索引(分类索引)使用。

3) 字顺表

《社会科学检索词表》的字顺表收入的全部主题词包括正式叙词和非正式叙词。叙词款目通常由款目叙词、分类号、参照项组成, 分类号用于揭示与分类体系的联系。非叙词款目由非叙词和用项参照组成, 一般不列出分类号。另外, 为了降低组配标引的难度, 词表还对组代词予以揭示。以下为叙词款目、非叙词款目和组代词的样例。

叙词款目样例:

拜占庭帝国
F132
J 395—1453 年
D 大秦
东罗马帝国
S 世界中世纪史
F 巴列奥洛王朝
拜占庭灭亡汪达尔
.....
C 拜占庭军区制
.....

非叙词款目样例:

东罗马帝国
Y 拜占庭帝国

组代词样例:

档案标引+主题标引

D 档案主题标引

档案主题标引

Y 档案标引+主题标引

4) 英汉译名对照表

英汉译名对照表是词表的辅助工具,主要供对外交流文献时英文译名对照参考使用。本部分共收录翻译名词 12 000 条,全表不标注分类号和参照符号。英文译名的排列以单词为单位,逐一按英文字母进行。英文译名多用名词和形容词,少用介词、连词、副词和数词。

5) 概要评价

《社会科学检索词表》作为我国社会科学领域的一部多学科词表,有如下主要特点。① 采用了分面叙词表的模式,这是分类语言与主题语言的最佳结合,其分面类表与主题词表两部分的类目与主题词原则上一一对应,比一般的分类主题结合方式更加精确,两者既是对方的索引,本身又是标引和检索的工具,通过两者的结合,可提高标引和检索的效率。② 分类表部分基本上采用了分面结构,在分类表的编制中重视传统分类与分面模式的结合,属于体系分类与分面分类相结合的半分面分类法,探索了适合我国文献标引使用的社会科学各领域的主题分面体系。③ 重视类表编制中学术性、实用性的结合,如分类表部分对传统分类体系的吸取,叙词的选择较为详尽,有较强的学术性和使用价值。

《社会科学检索词表》也存在一些不足,主要表现在如下几方面。① 主要分面的设置不够充分,不能涵盖所有的分面。各大类下的分面一般按照统一规定的分面次序排列,但也有一部分类目的分面与规定的统一次序不一致。② 分面叙词表的大类设置上有欠缺,如把《中图法》文化类中的新闻学、传播学、图书馆学、档案学、教育学分别作为三个大学科,却没有文化大类。③ 在选词上应增加一些比较关键的主题词,如中共党史类、人物研究类、儿童问题类等的类目和主题词。④ 一些用代关系欠妥,如苏联史 Y 俄国史、科技 Y 科学等。⑤ 虽然类表采用了分面结构,但是除了主表类目与通用类之间可以进行组配外,对主表类目之间的组配并没有规定,其标记符号也不适合组配,不能充分发挥其组配功能。在个别分类号的使用上也欠严谨、规范。

4. 《中国分类主题词表》

1) 概况

《中国分类主题词表》是在《中图法》类目与《汉表》主题词对应的基础上,将分类法与主题法融为一体的一种信息资源标引和检索的工具。该表由国家图书馆为首的全国 40 余家图书情报单位共同编制,编制工作从 1986 年发起到 1994 年出版,历时 8 年,先后有 160 位专家学者、专业人员参加。全表包括《分类号—主题词对应表》和《主题词—分类号对应表》两部分,共分 2 卷 6 册,收录分类法类目 50 317 个,主题词 101 376 个,主题词串 102 000 个,入口词 14 690 个,它是我国进行分类标引和主题标引的主要工具。

从 2000 年开始,国家图书馆中图法编委会对《中国分类主题词表》进行了修订,出版了《中国分类主题词表》(第二版),包括印刷版和电子版。

2) 结构

《中国分类主题词表》包括《分类号—主题词对应表》和《主题词—分类号对应表》两部分。

(1) 《分类号—主题词对应表》的结构。

《分类号—主题词对应表》以《中图法》和《资料法》的类目体系为基础,把《汉表》

主题词及主题词组配形式对应于各级类目之下编制而成，是从分类角度进行分类主题一体化标引的工具。《分类号—主题词对应表》的款目包括：分类号、类名、类目注释及对应的主题词、主题词串、对应参见和注释。对应款目的编排格式分为左右两栏，中间用竖线隔开。左栏为《中图法》和《资料法》的分类号、类名与注释；右栏为相对应的主题词或主题词的组配标题形式。当一个类目有多个对应主题时，各主题词之间用分号“；”隔开。例如：

H08	应用语言学	应用语言学
H085	机器翻译	机器翻译
	自动化翻译的理论著作入此， 翻译机入 TP391.2	单语种翻译；翻译规则系统（机 器翻译）机器语法（机器翻译）； 机助翻译；机器翻译—理论；词 对词翻译（机器翻译）；句对句翻 译（机器翻译）；译语（机器翻译）； 原语（机器翻译） 注：翻译机用 TP391.2 下对应 的主题词标引
H085.2	原文的自动分析与综合	原语（机器翻译）—分析；原语 （机器翻译）—综合； 一对一翻译（机器翻译）
H085.3	两种语言的翻译	

从上述样例看出，《分类号—主题词对应表》是以分类体系为中心展开的，每个分类号后，对应相应的主题词或主题词的组配形式，从而可以从学科分类着手查找相应类目及对应的主题词和主题词的组配标题。

《分类号—主题词对应表》中分类号与主题词的对应方法如下。

第一，类目与主题词的对应，以《中图法》、《资料法》的实有类目为基础，原有类目及其注释基本不做增删。

第二，类目对应的主题词中，首先列出与类名概念对应的主题词及主题词串（黑体），同时列出类目注释中的概念对应的主题词和类目包括的主题词。例如：

P 125.11	日食	日食
		日环食；日偏食；日全食

第三，类目对应的主题词必须是主题词表中相对应的正式主题词。如果没有相应的专指主题词，则可采用最接近的主题词的组配标题形式。例如：

B844.5	妇女心理学	应对应于 妇女心理学
Q13	生物形态学	应对应于 生物学：形态学

第四，《汉表》中每个主题词均根据其学科属性分类，归入相应的知识门类下。在主题词无确切类目对应时，一般采用上位类标引或靠类标引的方法，对应于相应类目下。例如：“战略学”一词可靠类对应应在“C934 决策学”类目下。

可与多个类目对应的主题词，应选择一个主要类目对应，并可在另一个或两个类目下加方括号列出（作为次要类号）。例如：“自我刺激”可对应在两个类目下，其中一个为主要类目，B842.6 情绪与情感；另一个为次要类目，[B845] 生理心理学。

第五，在使用多个主题词对一类目组配对应时，其引用次序应依据通用分面公式（国家标准《文献主题标引规则》），按“主体因素—通用因素—位置因素—时间因素—文献类型因素”的次序确定。例如：

G526.6	教育统计资料	教育统计—统计资料—中国
--------	--------	--------------

第六,对应款目中使用的符号。

加号(+),《资料法》与《中图法》类目体系一致,但详略程度不同,凡分类号中插有加号的号码,加号前的部分是《中图法》和《资料法》共同的号码,加号后的部分为《资料法》的类目延伸加细的号码,如 R122.1+1 化学分析,《中图法》的分类号是 R122.1;《资料法》的分类号是 R122.11。若加号位于类号之前,则单纯是《资料法》的类号,如+Q959.421 虎鲨目。如果不带加号,则表示两者的分类号相同。

冒号(:),用于对应表的右栏,是主题词概念交叉组配的符号。交叉组配是指选用若干个具有交叉关系的主题词进行组配,表达一个复合的概念。例如,“环境微生物学”可标引为:

环境科学:微生物学

横(—),用于对应表的右栏,是主题词概念限定组配的符号。限定组配是以表示事物的主题词和表示事物特称、属性、方面的主题词进行组配,表示一个新的专指概念。例如,“汽车发动机”可标引为:

汽车—发动机

逗号(,),用于对应表的右栏,表示倒置词序中两个主题词间组配的符号,或者用于对前面的主题词起修饰作用的自然语言语词。例如:

轨道(铁路),重型;调味品,常用Δ

方括号([]),用于对应表的左右栏,左栏表示交替类号,右栏表示交替类号对应的主题词,以及对应于多个类目的主题词(仅对在非主要类目处的主题词加[])。例如:

[R122.7] 大气污染及其防护	[空气污染]; [空气污染—防污防治]
C19 创造发明、先进经验	社会科学—创造发明; 社会科学—先进经验
	[创造心理学]; [创造学]

三角号(Δ)用于对应表的右栏,表示对主题词起修饰作用的自然语言语词。例如:

乙醇,食用Δ—酿造

分号(;)用于对应表的右栏,作为多个主题词或主题词串之间的分隔符号。

(2)《主题词—分类号对应表》的结构。

《主题词—分类号对应表》是以主题词的字顺排列为基础,把分类号对应于各个主题词或主题词串之下编制而成的,是从主题词角度查找主题词和分类号,进行分类主题一体化标引的工具。

其款目构成要素包括:主题词及参照项、主题词串及对应的分类号、各种符号和对应注释等。款目格式为:主题词在上,其下依次列出分类号、注释等。每个主题词下揭示其 Y(用)、D(代)、Z(族)、C(参)等语义关系,族首词下则对等级关系实行全显示,主题词串下则不显示其语义参照关系。

例 1:

辨识

N94

D 有认识能力的系统

- . 残差辨识
- . 离线辨识
- . 区域辨识
- . 系统辨识
- . 相关辨识

- . 在线辨识
- C 识别
- C 信号特征分析

例 2:

辨识—生物模拟

Q811.212

《主题词—分类号对应表》的条目基本上是在《分类号—主题词对应表》的基础上反向对应而成的,其词汇以《汉表》原有的主题词为依据,并进行了大量增补、删除和修改,以更加适应文献标引实践的需要。结合不同的情况,词表编制时主要采用了以下方法:

第一,当某一主题词或主题词组配形式对应多个分类号时,各分类号之间用分号分隔,例如:

物理有机化学

O621; O621.16

物理性质试验—建筑声学

TU112.2; TU112.2+ 4

第二,单个主题词,包括新增词和非正式主题词,均应显示词间关系。非正式主题词后只采用 Y 项,正式主题词后取消 S、F 参照,只采用 D、Z、C 项。当主题词为族首词时,在族首词下将其等级关系加以全显示。

第三,主题词串即主题词组配标题,主要以《分类号—主题词对应表》列出的组配标题为基础,其下给出分类号,组配标题不需要显示词间关系,但须作选择性轮排。

第四,规定复分表、组配因素的圈码,加在主表分类号后部或附表分类号之前,指示分类号、主题词的组配标引。为便于标引,对应表规定 9 种圈码的含义如下:

圈码① 一按“总论复分表”分;

圈码② 一按“世界地区表”分;

圈码③ 一按“中国地区表”分;

圈码④ 一按“国际时代表”分;

圈码⑤ 一按“中国时代表”分;

圈码⑥ 一按“中国民族表”分;

圈码⑦ 一按“专用复分表”分;

圈码⑧ 一该词为主题词串中的一个组配因素,不能独立使用,必须与其他主题词组配才能使用;

圈码⑨ 一按《资料法》中的“通用时间、地点复分表”分。

例 1: 对文献主题“中国四川人口调查”进行分类标引。

首先在《主题词—分类号对应表》中查得人口调查的分类号是: C924.25 ③,根据圈码③的提示,应利用“中国地区表”进行复分,因此该文献的分类号为 C924.257.1。

例 2: 对文献主题“植物药的临床应用”进行分类标引。

首先在《主题词—分类号对应表》中查得植物药的分类号是: R282.71 ⑦,根据圈码⑦的提示,应利用专用复分表进行复分,因此该文献的分类号为 R282.710.7。

3) 《中国分类主题词表》第二版

为了保持与 1999 年修订后的《中图法》分类一致,适应近年来新学科、新技术和社会迅猛发展所带来的新词汇的大量增加,以及词汇间语义关系更加复杂等情况,提高文献标引和检索的质量,《中图法》编委会决定从 2000 年开始对《中国分类主题词表》进行修订,于

2005年9月出版了《中国分类主题词表》第二版,包括电子版和印刷版。电子版为首次研制,其体系结构由一个主框架窗体和多个子窗体构成。主窗体通过自动显示滚动条来控制浏览所有子窗体。子窗体由3个文档构成,包括“分类号—主题词对应表”文档,窗口标题简称为“分类表”;“主题词—分类号对应表”文档,窗口标题简称为“主题表”;“词族表”文档,窗口标题简称为“词族表”。电子版还提供一个在形式上与印刷版基本相同但不属于子窗体形式的“浏览表”窗体,窗口标题简称为“浏览表”,包括“分类号—主题词对应表”和“主题词—分类号对应表”。

印刷版包括2卷6册。第1卷为《分类号—主题词对应表》(2册),第2卷为《主题词—分类号对应表》(4册)。第1卷与电子版内容完全相同;第2卷省略了电子版的部分内容,如主题词英译名、名称主题词(包括人名、团体机构名、题名),类目对应的组配标题(词串)。《分类号—主题词对应表》以《中图法》(第四版)为主体,将原《汉语主题词表》中的主题词及新增主题词对应于相应类目下。《主题词—分类号对应表》以原《汉语主题词表》的字顺表为主体,增加了大量的主题词串,将《中图法》的全部分类号对应于相应的主题词及主题词串下。

《中国分类主题词表》第二版是我国目前规模最大的分类主题一体化标引工具,共收录分类法类目52 992个、主题词110 837条、主题词串59 738条、入口词35 690条,包括哲学、社会科学和自然科学所有领域的学科和主题概念。可适用于图书馆、档案馆、情报所、书店、电子网站等进行各种类型、各种载体文献的分类主题一体化标引和检索。

《中国分类主题词表》第二版《分类号—主题词对应表》样例:

C933	领导学 领导哲学、领导心理学入此	领导学 领导;领导思维学;领导系统;领导心理学; 领导行为;领导行为理论;领导学\哲学;领导者
C933.1	领导体制 领导体制的原则和评价、各种领导体制、一长制、集体领导等入此	领导体制;领导体制\原则;领导体制\评价 一长制
C933.2	领导方法 领导艺术、领导思想、领导效能、各种领导类型、党政领导、行政领导等入此	领导方法;领导\类型;行政管理\领导 党政领导;领导思想;领导效能;领导艺术

《中国分类主题词表》第二版《主题词—分类号对应表》样例:

例1: xi qu

戏曲

E Chinese Opera

J82

D 戏曲艺术

. 古代戏曲

.. 参军戏

.. 传奇剧(戏曲)

.. 傀儡戏
.. 杂剧
.. 滑稽戏
.....

例 2: 镭矿床\金属矿开采

TD868

《中国分类主题词表》第二版中使用了如下符号。

(1) 保留分类号中间的加号 (+), 作为区别图书、资料使用的标记, 不用于标引文献;

保留方括号 ([]), 只用于表示交替类目及其对应的主题词;

保留分号 (;), 作为多个主题词或主题词串之间的分隔符号。

其他第一版中使用的符号, 如冒号 (:)、短横 (-)、逗号 (,)、三角号 “Δ” 在第二版中均不再使用。

(2) 增加斜杠符号 “\”, 用于表示概念相交或概念限定或倒置关系的主题词之间的组配。

例如:

“月掩恒星” 标引为 “月掩星\恒星”

“造血机能” 标引为 “造血系统\机能”

“各国军事地理” 标引为 “军事地理\各国”

(3) 增加双竖线符号 “||”, 当一个主题词对应多个类目时, 作为对应非主要类目的指示。

例如:

R543 血管疾病

雷诺氏 (Raynaud) 患者

入 R747.3; 红斑性肢痛症入 R747.4

血管疾病

|局部缺血; 库欣综合征;

血栓栓塞; 血栓形成

(4) 参照符号和圈码的使用与第一版相同。

4) 《中国分类主题词表》评价

(1) 通过将《中图法》类目与《汉表》主题词的对应, 建立起一个分类语言与主题语言结合的一体化工具, 可以利用它同时进行分类主题的标引和检索, 简化了操作程序, 降低了标引难度, 改进了标引和检索的质量。

(2) 其分类部分是将《中图法》、《资料法》融为一体的类目体系, 可以同时供图书资料单位标引使用。

(3) 其主题法部分除收有原有的叙词外, 还包括近年来中文图书标引中新增的叙词和对应表编制时的新增词, 以及《分类号—主题词对应表》中出现的主题词组配形式, 是《汉表》叙词比较完整的版本。

(4) 《中国分类主题词表》第二版推出了电子版, 免除了用户在印刷版的六大分册之间来回翻检的烦恼, 增强了词表的易用性, 提高了标引和检索的效率。《中国分类主题词表》第二版的整体性能得到了提升, 主要表现在: 扩充了词表的规模; 改善了词表的性能; 简化了词表中的各种专用符号, 电子版和印刷版进行了必要的分工, 使用更加便捷。

虽然《中国分类主题词表》第二版已经在第一版的基础上做了大量的修订, 但仍然存在一些不足, 主要表现在以下几个方面: ① 类目对应标引深度较低, 不能满足自动分类的需要; ② 词表修订周期过长, 词汇更新滞后; ③ 词表的性能和功能有待完善; ④ 印刷版版面设计有待改进。




本章小结

无论是传统的代码语言、分类语言、主题语言，还是新出现的本体语言，它们都是对信息外部特征和内容特征进行描述和揭示的语言系统。随着信息检索技术、方式、手段的变革，信息描述语言也经历非控→先控→后控的不同发展阶段，但无论信息描述语言怎样发展，“词汇控制是永远不会消失的，变化的只是词汇控制的方式、方法和手段。”



问题讨论

1. 网络环境下，自然语言是否会取代规范语言？
2. 随着全文检索的普及，信息标引是消亡还是更加深化？
3. 本体与传统情报检索语言之间有何异同？




第4章

信息著录法

本章引言

将信息实体的有关特征著录下来,就是信息著录。有关如何著录的具体规定和具体做法就是信息著录法。按照一定的著录法对某一信息实体的内容和形式特征所做的描绘就形成一条款目或记录。一条款目或记录是一种信息实体的高度概括和浓缩。有了款目或记录,才能对款目或记录进行分类、主题等各种标引,才能在标引的基础上对款目或记录进行编排,才能通过款目或记录对庞杂的信息资源进行有序的组织和控制。因此,著录出高质量的款目或记录是信息组织的基础性工作。要做好著录工作,必须了解各种著录标准和规则。本章分别介绍了传统著录法、机读目录著录法和元数据著录法。

本章重点

- 著录的含义与特点;
 - 传统著录法的著录项目及标识符;
 - 《国际标准书目著录》(ISBD);
 - 《英美编目条例》(第二版)(AACR2);
 - 《文献著录总则》;
 - MARC 记录的发展沿革和基本格式;
 - 元数据的内涵及功能;
 - DC 元数据的发展和元素定义;
 - MODS 元素组成;
 - CDWA 和 FGDC 元数据标准。
- 

4.1 传统著录法

传统著录法是相对于 MARC（机读目录）著录法和元数据著录法而言的，是指按照《国际标准书目著录》（ISBD）等标准和规则的要求统一进行规范化的著录，但著录方式是手工操作或利用计算机进行分项著录，形成规范的卡片目录，然后由手工编排各款目著录法。传统著录法的原理对理解 MARC 和元数据著录均有助益。

4.1.1 传统著录法概述

1. 什么是著录

“著录”一词在我国具有悠久的历史。它的原意是指在簿籍上的记载，后来用以泛指在任何载体上的记载。在国外，著录的英文对应词为 Description、Descriptive 或 Bibliographical Description（书目著录），这些英文词是国际编目界通用的术语。但长期以来，作为专业术语的“著录”的明确概念在我国一直没有形成。直到 1983 年《文献著录总则》（GB 3792.1—83）正式颁布，才给著录下了一个明确的定义，即：著录是指在编制文献目录时，对文献内容和形式特征进行分析、选择和记录的过程。

《中国文献编目规则》中关于“著录”的定义也与此大致相同，即：编制文献目录时，按照一定的规则对文献的形式特征和内容特征进行分析、选择和记录的方法和过程。

从上述定义中，可以看出“著录”主要包括以下两层含义。

（1）明确了著录的对象和内容。著录的对象是文献，著录的内容是文献的形式特征与内容特征。文献的内容特征是指文献的知识内容，如分类号、主题词等。文献的形式特征是指文献的实体形式，主要包括两个部分：一是关于文献外表的文字记载，如题名、责任者、版本、出版社、出版年、丛编名、标准文献号、价格等；二是关于文献物质形式的文字记载，如装订、尺寸、数量、图表等。

（2）明确了著录的基本方法与工作环节，即分析文献信息的特征、从中选择具有著录价值的内容、记录必要的文献目录信息。

著录首先要分析著录对象的特征，这是文献信息著录中的重要任务。在分析的基础上，再正确地选择著录内容，并客观、准确地记录下来，使目录能够揭示文献信息最本质的特征及其相互关系，满足用户的基本需求。

另外，该定义还明确指出，“著录”必须遵循一定的规则。著录工作依据的规则比较多，一般包括文献信息著录规则、文献信息标引规则、规范记录著录规则等。

著录是揭示文献信息特征及有关信息的有效方法，也是编制文献目录的基本方法。著录的结果可以客观地反映文献信息的概况，为文献信息的管理、检索与利用创造条件。

文献信息编目工作主要可以分为两大步骤：第一步是文献信息著录，第二步是目录组织。前者为后者提供基础，后者以前者为前提。换句话说，著录是编目工作的基础，目录的质量在很大程度上由著录的质量决定。虽然编目工作实现计算机化后，目录组织的很多工作可以由计算机自动完成，但著录工作当前仍然基本上由人工进行，而且，为了更加客观、准确、完整地反映文献，著录变得更加复杂多样。从这个意义上说，著录的重要性大大增加了。

2. 款目与记录

款目与记录是著录的结果，是目录编制的基本单元。

1) 款目

款目是指依据一定的规则和方法,对文献特征与编目业务信息所做的记录。其表现形式是反映文献内容特征和形式特征的著录项目的组合,包括描述项目、检索点(标目)、编目业务注记三部分信息。款目是组成传统目录的基本要素,其主要作用是:

(1) 揭示文献的检索点(即著录标目),明确各条款目在目录中的排列位置,提供检索途径;

(2) 揭示文献的主要形式特征与内容特征,提供认识、选择文献的依据;

(3) 揭示编目业务注记,以提供文献索取、管理及款目更新、管理的依据。

描述项目是指著录项目中用于揭示文献信息基本特征的事项,包括题名与责任说明项、版本项、文献特殊细节项、出版发行项、载体形态项、丛编项、附注项、文献标准编号与获得方式项。

描述项目的主要作用是:

(1) 客观描述文献信息特征;

(2) 概略反映文献全貌;

(3) 提供识别与选择文献的主要依据。

描述项目是各类型款目或记录的基础,也是著录内容的主体信息,其质量直接影响文献信息的识别与选择。

检索点是指用于目录记录或款目排序与检索标识的数据单元,包括标目与排检项或根查项的著录内容。标目通常位于款目的最上方,排检项或根查项则在款目的最后部位。检索点与标目是著录内容的重要信息,其质量直接影响目录的检索效果。

编目业务注记也称为图书馆业务注记,是指编目机构为了业务工作的需要而在款目或记录中所做的一些记载。这些记载通常都是特定的符号或略语,它们不描述文献信息的特征,而是反映文献信息存储或编目等方面的信息。

编目业务注记一般由具体编目部门设计与著录。传统的编目业务注记主要包括索取号、财产登记号、储藏地点记载、根查、注销登记等内容。机读目录中的这类编目业务注记内容更丰富,如头标区中的记录状态、编目等级,以及数据区中的国际使用字段、国内使用字段等,其作用主要是用于机构内部索取文献和管理文献,或者是作为更新、管理款目或记录的依据。不同的编目机构的具体情况不同,会使他们的编目业务注记在品种、数量及表现形式上存在较大的差别。

编目业务注记是文献著录信息的重要组成部分,既是目录管理与利用的重要依据,也是文献信息管理与利用的必要依据。随着该信息类型的不断增长,它的功能越来越强,在当代著录信息中具有重要的作用。

2) 记录

《中国文献编目规则(第二版)》给“记录”下的定义是:记录是指表述事物的特征,具有完整的含义,从内容和使用的角度能作为一个整体来识别的一组相关数据项的组合。

在编目领域,通常将以机读形式存储于目录数据库中的目录数据称为“记录”。

一条记录相当于手工编目中的一条“款目”,但记录所“著录”的内容更为丰富、复杂,不仅极大地扩充了“款目”上的信息,还增加了代码信息及计算机识别与处理的符号。

3. 著录信息源

著录信息源(Source of Information)是指款目或记录中著录信息的来源。明确著录信息源是准确、一致地进行文献著录的保障。

著录信息源可以划分为主要信息源与参考信息源。另外,还有更为具体的规定信息源。

1) 主要信息源

主要信息源是指在著录中优先选作著录信息来源的文献信息组成部分。

著录信息的基本来源是文献信息本身，是被著录的整部文献信息。各类型文献信息的著录信息源，均为被著录的文献信息本身。例如，普通图书的有关形式特征的著录信息主要来源于题名页、版权页、封面、书脊、附录等处，有关内容特征的著录信息则主要来源于正文、书名、目次、序跋文字和内容提要等。当然，同一文献信息的不同组成部分所反映的同一文献特征也有不同的情况，如有的图书的封面或书脊上的题名与书名页上的题名不同。因此，为了保证著录的一致，就必须进一步明确著录信息在文献中的具体来源部分，还要确定不同著录信息源选用的优先顺序。而主要信息源就是著录规则中明确规定的，在著录中必须依次选用的文献信息组成部分。

各种类型文献信息均有其各自特定的主要信息源。AACR 2 中各类型文献信息的主要信息源如表 4.1 所示。

表 4.1 AACR 2 中各类型文献信息的主要信息源

文献信息类型	主要信息源
图书、小册子与散页出版物	题名页
测绘制图资料	a) 资料本身 b) 容器或函套、球仪的支架与底座等
手稿	手稿本身、题名页、书尾题署、手稿头等
乐谱	题名页、封面、卷首
录音资料	资料本身、容器与标签
影片与录像资料	a) 资料本身 b) 容器
图示资料	资料本身、标签、容器、附带文字资料等
电子资源	资源本身、题名屏幕、主选择单、程序说明、首先显示的信息、主页等
立体工艺品与实物	文献本身、标签、附带文字资料和容器
缩微资料	题名帧
连续资源	a) 题名页或封面 b) 物理载体上的标签

2) 参考信息源

参考信息源是指在著录中参考使用的信息来源，如有关工具文献与参考文献等。

之所以使用参考信息源，是由于主要来自于文献信息本身的著录信息源有时无法提供足够的著录信息，如文献残缺、特征不详或有误等。这时，就可以考虑使用文献信息本身之外的信息即参考信息源，利用各种工具文献与参考文献来解决著录中的问题，弥补主要信息源的不足。

3) 规定信息源

规定信息源是指各个著录项目及其单元著录信息的特定来源。为确保著录信息选取的一致性，著录规则还进一步规定了每个著录项目的著录信息源，通常为文献信息的某一个或某几个组成部分。著录各个著录项目及其单元时，必须依据规定信息源规定的内容及其先后顺序来选择使用信息源。各类型文献信息著录中，规定信息源也不尽相同。

4. 著录项目及著录用标识符

1) 著录项目

著录项目是指用以揭示文献信息形式特征与内容特征的记录事项，如题名与责任说明

项、出版发行项等。著录项目可以包括著录单元。著录单元是指著录项目的组成部分,如题名与责任说明项中的题名、责任者名等。

著录项目是依据文献信息本身的形式特征与内容特征,结合读者检索利用的特点与规律确定的,主要包括题名与责任说明项、版本项、文献特殊细节项、出版发行项、载体形态项、丛编项、附注项、文献标准编号与获得方式项。按照著录项目的重要程度,又可以将其划分为主要项目与选择项目。主要项目是著录中必不可少的,选择项目则可以根据具体情况决定是否选用。

著录规则中对著录项目的选择使用的详略做了明确的规定,即规定了著录的详简级次。

所谓著录的详简级次,是指根据著录项目或著录单元的详简程度划分的等级区别。例如,在 AACR 2 中,分为简要级次、基本级次与详细级次三个级次。划分著录详简级次的目的是为了能够更好地适应不同类型文献信息工作部门的实际情况与需要,兼顾各类型文献信息著录的特殊性,使同一个目录系统能够兼容详简不同的著录款目。

著录项目的设置一般应该具有规定性、兼容性和灵活性。具体来说,规定性主要表现在对著录项目的名称、数量、各项的排列顺序等方面有明确的规定,使其具有相对的稳定性;兼容性主要表现在著录项目基本概括了各类型、各种具体文献的共性,反映了不同文献的特征;灵活性主要表现在著录项目的应用方面,各编目机构可以根据文献信息的特征与本机构的具体要求选择使用项目,进而编制详略程度不同的款目。

2) 著录用标识符

著录用标识符是指著录中用以识别著录项目及其单元的特定符号。一般包括著录项目标识符号与著录单元标识符号。著录项目标识符统一置于著录项目之前,如“.--南京”中的符号“.--”。著录单元标识符则置于著录单元之前或外部,如“:江苏教育出版社”、“(精装)”中的符号“:”与“()”。

在文献信息著录中采用标准的著录用标识符的主要作用是,能跨越不同的语言文字障碍,实现国际文献目录信息的交流与共享。

4.1.2 文献信息著录规则

文献信息著录规则是指根据文献信息本身的客观情况,结合读者检索要求而制定的一整套系统记录文献信息特征的原则和方法。

著录规则是编目工作发展到一定时期的产物,是人们从长期编目工作实践中总结出来的基本原则和规律,也是编目工作制度化、规范化的结果。文献信息著录规则的主要作用是:指导文献信息著录工作,处理文献信息著录中的一般性问题,使文献信息著录保持一致性,使各具特色的文献信息在目录中有相对统一的表现形式。

科学实用的文献信息著录规则是文献编目工作的前提。而且,文献信息著录规则还要随着文献信息及读者检索要求的变化而逐渐更新,不断提高其标准化程度。

著录规则是编目工作者的操作规程,无论是手工编目还是计算机编目,都必须遵守著录规则。只有严格遵循规则,才能确保目录的质量。因此,在编目之前,必须明确所使用的有关著录规则。

文献信息著录规则经历了数百年不断改进、不断完善的发展历程。按照不同的标准,可将现有的规则划分为不同的类型:按照著录规则所涉及的编目内容范围,可将其划分为描述著录规则、包括描述与标目著录的规则、包括描述与标目著录及目录组织的规则;按照著录规则选择著录对象的范围,可将其划分为单一著录对象的著录规则、多种著录对象的著录规则;按照著录规则的表述形式,可将其划分为将总则与各分则分别编制的著录规则、将总则

与各分则融于一体的著录规则。现行的著录规则，其结构与内容一般包括引言、描述著录、标目法、附录、名词术语等部分。

下面简单介绍国内外主要的几个著录规则。

1. 《国际标准书目著录》

《国际标准书目著录》(ISBD, International Standard Bibliographic Description) 由 ISBD 修订委员会推荐, IFLA 编目专业组常设委员会通过。自 1971 年以来, 先后出版了 ISBD (G) 与一系列著录各种不同类型文献的 ISBD 分则, 而且从 1978 年开始, 其版本不断修订更新。

1) ISBD 的发展

1969 年, 国际编目专家会议建议成立“ISBD 工作组”。1971 年年底, 由 IFLA 设立的 ISBD 工作组编制出了《国际标准书目著录(专著)》(ISBD (M)) 初稿。该稿一经发布就得到了英国、原联邦德国、法国、加拿大、澳大利亚等国的赞同与应用。1974 年, 在伦敦正式出版了该条例的修订本——“第一标准版”。

在 ISBD (M) 成功的基础上, IFLA 又相继成立了几个分则工作组, 编制 ISBD (S)、ISBD (NBM)、ISBD (CM) 等分则。这些分则在编制的过程中, 出现了与 ISBD (M) 的内容不尽一致的问题。1975 年 10 月, IFLA 在巴黎召开会议, 讨论制定一个总的框架, 作为各分则编制的指南, 以达到协调各个分则的目的。1976 年, ISBD (G) 初稿完成, 并提交 IFLA 编目专业组常设委员会组织讨论。1977 年, 正式出版 ISBD (G) 第一标准版。

ISBD (G) 出版之后, 新编制的分则都必须以此为编制依据, 已经出版的分则也根据它进行了修订。这样, 以 ISBD (G) 为总原则、针对不同类型的文献而设计的一系列分则陆续出版, 形成了一整套国际标准的著录规则。

从 1978 年开始, ISBD 已有的一系列分则广泛征求各方意见, 不断进行必要的修订, 使其在内容上更加符合编目工作的具体要求, 在形式与结构上更加规范统一。2006 年国际图联 (IFLA) 提出了 ISBD 统一版。2007 年正式出版了 ISBD 初级统一版。

2) ISBD 的编制体例

与以往的著录规则相比, ISBD 在编制体例上有所创新。ISBD 的总则与各个分则分别制定、陆续出版, 是一套既紧密联系, 又相对独立、自成体系的著录规则。

具体来说, 在一整套 ISBD 规则中, ISBD (G) 为各种 ISBD 分则的制定提供了框架, 起到了控制的作用, 但不用来对任何具体类型文献进行信息著录。每一种分则都既是总则编制原则具体化的产物, 又各自独立, 能系统地解决特定类型文献信息著录问题。整套规则中, 总则与各分则、各个分则之间都互相联系、互相参见, 构成一个有机的整体。

目前, 由 IFLA 颁布的 ISBD 系列规则有:

《国际标准书目著录(总则)》[ISBD (G), General International Standard Bibliographic Description], 1977 年第 1 标准版, 1987 年第 2 版;

《国际标准书目著录(专著)》[ISBD (M), International Standard Bibliographic Description for Monographic Publications], 1971 年推荐本公布, 1974 年第 1 标准版, 1988 年第 2 版, 2001 年修订版;

《国际标准书目著录(连续出版物)》[ISBD (S), International Standard Bibliographic Description for Serials], 1977 年第 1 标准版, 1987 年第 2 版;

《国际标准书目著录(地图资料)》[ISBD (CM), International Standard Bibliographic Description for Cartographic Materials], 1977 年第 1 标准版, 1987 年第 2 版;

《国际标准书目著录(非书资料)》[ISBD (NBM), International Standard Bibliographic Description for Non-Book Materials], 1977 年公布, 1987 年第 2 版;

《国际标准书目著录（古籍）》[ISBD (A), International Standard Bibliographic Description for Antiquarian Materials], 用来著录 1801 年以前出版的专著, 1980 年公布;

《国际标准书目著录（乐谱）》[ISBD (PM), International Standard Bibliographic Description for Printed Music], 1980 年出版, 1987 年第 2 版;

《国际标准书目著录（计算机文件）》[ISBD (CF), International Standard Bibliographic Description for Computer Files], 1990 年出版;

《国际标准书目著录（电子资源）》[ISBD (ER), International Standard Bibliographic Description for Electronic Resources], 1997 年出版, 由 ISBD (CF) 修订而成, ISBD (CF) 修订小组推荐, ISBD (ER) 将电子资源分成本地电子资源和远程检索电子资源（即网络信息资源）;

《国际标准书目著录（分析著录）》[ISBD (CP), International Standard Bibliographic Description for Component Parts], 1982 年出版;

《国际标准书目著录（连续出版物与其他连续资源）》[ISBD (CR), International Standard Bibliographic Description for Serials and Other Continuing Resources], 2002 年出版。

3) ISBD 的编制结构与内容

ISBD 总则与各个分则的编制结构和内容基本相同, 都由以下三个部分组成。

(1) 概述 (Preliminary): 该部分阐明本条例的编制范围、目的与应用, 以及有关定义、信息源、著录与语言文字等。

(2) 著录单元说明 (Specification of Elements): 该部分内容最多, 详细地说明了每个著录项目及其著录单元的著录规则。

(3) 附录 (Appendices): 该部分包括多层次著录、双向行文记录、具体的著录样例等。

从内容上看, ISBD 主要是针对文献信息的描述规则所做的具体规定, 目的是解决文献信息描述的标准化问题, 但没有涉及标目的选取与著录等规则。

4) ISBD 的主要特点

ISBD 是一整套文献信息著录的国际标准, 它具有如下主要特点。

(1) 编制目的明确。

ISBD 编制的总体目的是: 促进国际书目信息交流, 实现文献信息资源共享。具体目的是: 使各国的书目著录具有互换性, 各国的书目记录易于识别, 传统的手工目录易于转换为机读目录。

(2) 措施具体有效。

ISBD 具体规定了著录项目、著录项目的顺序及著录标识符。所采取的这些措施具体而有效。表 4.2 是 ISBD (G) 的著录项目及其标识符。

表 4.2 ISBD(G)的著录项目及其标识符

Area (著录项目)	Prescribed Preceding (or enclosing) punctuation for Element (标识符)	Element (著录单元)
Note: Each area, other than the first, is preceded by a point, space, dash, space (. —)		
1. Title and statement of responsibility area	[]	1.1 Title proper
	=	1.2 General material designation
	:	1.3 Parallel title
	:	1.4 Other title information
	/	1.5 Statements of responsibility
	:	First statement Subsequent statement

续表

2. Edition area	= / ; ,	2.1 Edition statement 2.2 Parallel edition statement 2.3 Statements of responsibility relating to the edition First statement Subsequent statement 2.4 Addition edition statement 2.5 Statements of responsibility following an addition edition statement First statement Subsequent statement
3. Material (or type of publication) specific area		
4. Publication, distribution, etc., area	; : [] , (: ,)	4.1 Place of publication, distribution, etc. First place Subsequent place 4.2 Name of publisher, distributor, etc. 4.3 Statement of function of distributor 4.4 Data of publication, distribution, etc. 4.5 Place of manufacture 4.6 Name of manufacture 4.7 Data of manufacture
5. Physical description area	: : +	5.1 Specific material designation and extent of item 5.2 Other physical details 5.3 Dimensions of item 5.4 Accompanying material statement
6. Series area	= : / ; , ;	6.1 Title proper of series or sub-series 6.2 Parallel title of series or sub-series 6.3 Other title information of series or sub-series 6.4 Statements of responsibility relating to the series or sub-series First statement Subsequent statement 6.5 International Standard Serial Number of series of sub-series 6.6 Numbering within series or sub-series
7. Note area		
8. Standard Number (or alternative) and terms of availability	= : ()	8.1 Standard Number (or alternative) 8.2 Key title 8.3 Terms of availability and/or price 8.4 Qualification (in varying positions)

可以看出, ISBD (G) 共规定了 8 个著录项目, 而其中的 6 项又进一步划分出若干个著录单元。在 ISBD (G) 这个总则中, 没有标明著录项目与著录单元的重复或选择使用问题, 这些是各个具体的 ISBD 分则的内容。

(3) 适用范围广。

作为国际标准, ISBD 对于不同地区、不同语言、不同规模、不同类型的文献机构都具有通用性。

ISBD 的制定和出版发行具有重大意义, 影响极其深远, 在统一各国文献信息著录条例、实现目录著录国际标准化方面做出了重要贡献, 在亚洲、非洲、欧洲、美洲的一些国家的书目中得到了广泛的应用, 不仅如此, 许多国家还将其作为制定或修改各种编目条例的重要参考依据。

2. 《英美编目条例》(第二版)

《英美编目条例》(第二版) (AACR 2, Anglo-American Cataloging Rules 2nd ed.) 由美国图书馆协会、英国图书馆协会、加拿大图书馆编目委员会、英国图书馆、美国国会图书馆联合提出, 由戈尔曼 (Michael Gorman) 与温克勒 (Winkler, Paul W.) 负责编辑, 于 1978 年在芝加哥、伦敦与渥太华同时出版。1988 年, 出版了修订版 AACR 2R, 1988。1998 年出版新的修订版 AACR 2, 1998 Revision。2002 年又出版了新的修订版 AACR 2-2002。

1) AACR 2 的发展

ISBD 的问世和推广, 在解决目录描述著录规则标准化问题方面取得了很大的成功, 这就更加迫切地需要编制出与其相匹配的、完整的著录规则。与此同时, 计算机在编目领域的广泛应用及非书资料的大量涌现, 使得原有的《英美编目条例》(AACR) 日趋老化的客观事实变得越来越突出, 因此, 条例的更新势在必行。

实际上, 美、英、加三国从 1974 年就开始了合作修订 AACR 的工作。经过几年的努力, 于 1978 年出版 AACR 2。这次修订的主要成就是, 将主要用于卡片式目录的传统条例发展为适应机读目录编制与多样化检索的条例, 顺应了编目事业的发展趋势。因此, AACR 2 被翻译成多种文字, 为许多国家的图书馆所采用。

社会经济与科技的发展, 网络技术的普遍应用, 文献信息资源的变化, 以及人们信息控制需求的加强, 都要求 AACR 2 不断发展和完善, 因此, AACR 2 只有不断修订再版才能满足编目事业的需要。

1983 年, 修订 AACR 2 联合指导委员会 (JSC, Joint Steering Committee for Revision of AACR) 明确地提出了修订 AACR 2 的方针和步骤。1986 年, 在原有编制成员的基础上, 将澳大利亚编目委员会增补为新成员。1988 年, 正式出版 AACR 2R。AACR 2R 的主要特点是, 吸收了过去十年的修订成果, 对一些内容进行了适当的调整, 增强了“计算机文档”的著录规则。1993 年和 1998 年, AACR 2 又相继进行了一些重要的修订, 并出版了修订版。

2002 年, 面对编目工作中出现的新问题, 再次对 AACR 2 进行了修订并出版新版 AACR 2-2002。AACR 2-2002 继承了此前各个版本的精华, 融合了 1999 年、2001 年修订本的内容, 在此基础上增加了 2002 年批准确定的新条款。该版修订突出重点, 如主要修订了总则、测绘资料、电子资源、连续出版物和组合资源 (Integrating Resources)、选择检索点的著录规则。不仅如此, 该版在对编目理论的发展方面也取得了一定的成果, 如重新界定与划分了书目资源类型, 更新了有关编目的一些基本概念等, 使得编目条例与书目资源种类的发展相互适应。凡此种种, 可以说, AACR 2-2002 修订版的产生, 不仅对当时现实的编目工作具有较强的适应性, 而且对编目理论和实践在未来的进一步发展和完善也具有开创性的意义。

从某种意义上说, AACR 2 是国际编目界共同努力的成果。美国、英国、加拿大、澳大

利亚等国家的许多编目专家不断为 AACR 2 的修订提出建设性的方案, 美国图书馆协会 (ALA) 及其下属机构、国际图联 (IFLA)、OCLC 等机构更是定期组织会议, 讨论 AACR 2 现存的问题与未来的发展。尤其是 JSC, 作为维护 AACR 2 的专门机构, 切实履行着其职责: 支持有效的书目编目实践, 维护与发展 AACR 已经确立的书目原则, 及时修订编目规则以反映用户需求和信息环境的变化。可以说, 在 AACR 2 的修订过程中, 它一直起着核心作用, 使得 AACR 2 不断修订, 与时俱进。

2) AACR 2 的编制结构与内容

AACR 2 是一部适用于多种类型、多种文字、多种载体的文献信息著录条例。以 2002 年修订版为例, 它的结构如下:

Part I Description (第一部分 描述著录)

- 1 General Rules for Description (著录总则)
- 2 Books, Pamphlets, and Printed Sheets (图书、小册子、散页出版物)
- 3 Cartographic Materials (测绘制图资料)
- 4 Manuscripts (Including Manuscript Collections) (手稿)
- 5 Music (乐谱)
- 6 Sound Recordings (录音资料)
- 7 Motion Pictures and Video recordings (影片与录像资料)
- 8 Graphic Materials (图示资料)
- 9 Electronic Resources (电子资源)
- 10 Three-Dimensional Artefacts and Realia (立体工艺品与实物)
- 11 Microforms (缩微品)
- 12 Continuing Resources (连续资源)
- 13 Analysis (分析)

Part II Headings, Uniform Titles, and References (第二部分 标目、统一题名与参照)

- 21 Choice of Access Point (检索点的选择)
- 22 Headings for Persons (个人著者标目)
- 23 Geographic Names (地理名称)
- 24 Headings for Corporate Bodies (团体名称标目)
- 25 Uniform Titles (统一题名)
- 26 References (参照)

Appendices (附录)

- A Capitalization (大写)
- B Abbreviations (缩写)
- C Numerals (数字)
- D Glossary (词汇)
- E Initial Articles (首冠词)
- Index (索引)

从以上结构可以看出, AACR 2 全书分为描述著录, 标目、统一题名与参照和附录三大部分。其中, 正文分为两大部分, 共有 19 章。这两个部分不分主次、互为补充, 形成了一部完整的著录条例。另外, 书末还有附录部分。

(1) 描述著录部分。

第一部分共有 13 章。第一章为著录总则, 阐明了著录的通用规则, 是以后所有各章著录规则的基础; 第 2~12 章是根据总则的基本原则, 针对特定类型文献信息制定的著录规则;

第13章则是部分文献信息通用的分析著录规则。在对具体类型文献进行著录时,应该选择使用与其相对应的分则。如果被著录的文献的类型比较复杂,则可以考虑使用多种相关的分则。例如,被著录的文献是连续出版的录音资料时,就既要用到第12章“Continuing Resources (连续资源)”的内容,又要用到第6章“Sound Recordings (录音资料)”的内容。

该部分的所有规则均以ISBD为依据。AACR 2完全采用了ISBD的8个著录项目及其顺序、著录标识符,并规定了著录用文字、著录信息源及各项目著录细则等。此外,还首创了明确款目详简内容的条款,列出了三个推荐的著录级次。

(2) 标目、统一题名与参照部分。

第二部分共有6章,其篇幅与著录部分非常接近。该部分的第21章规定了主要款目与附加款目检索点的选择条件;第22~25章规定了个人、地理、团体名称标目的形式与统一题名;第26章是编制个人与机构名称、地理名称和统一题名等参照的规则。在每一章中,将一般规则列在专门规则之前,当一个特定的问题没有特定的规则时,则可采用一般规则。该部分适用于著录各种文献信息的检索点,包括卡片式目录的主要款目标目与根查项的内容。

(3) 附录部分。

第三部分包括5种附录和一个综合索引。附录是AACR 2的重要组成部分,规定了著录使用文字的规范条款,是著录中不可缺少的内容。

3) AACR 2 的主要特点

(1) 贯彻了编目标准化的原则。

AACR 2十分重视编目标准化问题,关注编目标准的国际化与书目共享的发展趋势,注意保持与其他相关书目控制标准的协调性,特别是与国际标准书目著录(ISBD)、国际标准刊号(ISSN)及国际标准化组织(ISO)的标准保持一致。例如,在“描述著录”部分,AACR 2尽量与ISBD的发展保持一致,不仅保证了款目内容的统一著录,而且实现了不同类型文献信息的统一著录;在“标目、统一题名与参照”部分,AACR 2积极采用有关国际协议,全面吸收了1961年巴黎会议的标目原则,为其他国家标目法的制定提供了范例与样板。与此同时,AACR2的发展和完善也极大地促进了其他国际标准的更新。

(2) 改革了条例的组织结构与方法。

组织结构方面,AACR 2将“描述著录”列为第一部分内容,而将“标目、统一题名与参照”列为第二部分内容,颠倒了AACR等条例的组织结构的先后顺序,客观反映了编目程序的变化:旧编目工作程序是从标目法开始,由标目决定整个款目编制的方法;新编目过程是先描述著录,再选择检索点。

组织方法方面,AACR 2在对规则中各项条款的安排上突出了伸缩性的特点。例如,在章节的设置上有意识地留有扩展的余地:第一部分共有13章,第二部分却从第21章开始,将第14章到第20章作为空位,为条例内容的扩充留有充分的余地。

另外,对“描述著录”部分的条款所配编号系统也有规律可循,即凡内容性质相同的条款,均采用类似的编码。这就使得各章节条款与第1章的相对应,方便记忆与查阅使用。

(3) 沿用并更新了“主要款目”标目概念与“著者原则”。

AACR 2在沿用了西方传统编目条例的“主要款目”标目概念与“著者原则”,即将款目标目分为主要款目标目和附加款目标目并且主要以著者为主要款目标目的同时,也做了如下一些改进。

重新界定“著者”概念。强调只有对著作的知识或艺术内容负主要责任的个人与团体,才能被称做“著者”或责任者,缩小了著者概念的范围。因此,属于编辑、编纂的著作一律被改用题名作为标目,相对增加了题名作为主要款目标目的机会,削弱了著者主要款目的地位。

对责任者统一标目的规定更具有通俗性。强调选取个人或团体最为人们所熟知的名称及

其常见形式作为统一标目，改变了传统条例中对个人名称取原名、真名、全名，团体名称取官方正式名称的做法，增强了标目的实用性。

扩大了统一题名的范围。强调以不同形式、不同题名出版的同一著作均可使用统一题名来加以集中，改变了传统条例中仅对宗教经典、古代佚名著作和音乐资料使用统一题名的做法，更加突出了统一题名的作用。

（4）增强了条例使用的灵活性。

基于不同类型编目机构对编目的要求不同，AACR 2 非常注重条例在应用方面的灵活性。例如，提供多种文献类型的著录规则，便于不同编目机构选择使用；规定三个著录级次，满足不同编目机构的要求；确定一些机动性规则如交替性规则（Alternative Rules）、选择性规则（Optional Addition 或 Optionally）等，并注明“If appropriate”（如果合适）、“If necessary”（如果必要）等选择性短语，编目员可以根据特定情况灵活把握与使用。

AACR 2 在提供完备规则的同时，也留给编目员不少分析判断的自主权。这一特点在连续资源的著录中最为突出。

AACR 2 被公认为西方编目理论和实践的集大成之作，是国际合作修订编目规则的成功代表。它在兼顾传统编目特点的同时，适应了编目工作标准化、自动化、网络化的发展趋势，既为英语世界的文献信息编目工作提供了标准化的工具，也为各国编目规则的制定树立了典范，有力地推动了世界编目事业的发展。

AACR 2 已经成为真正意义上的国际编目条例，被许多国家接受、使用或者借鉴，如丹麦、芬兰、意大利、挪威、葡萄牙、西班牙、瑞典、土耳其等。另外，AACR 2 还有日文、中文、阿拉伯文等多种文字的版本，可见其影响之大、应用范围之广。

但 AACR 2 也存在一些问题，主要有：标目中仍然保留了“主要款目标目”与“附加款目标目”的概念，选取主要款目标目的规则复杂而烦琐，给使用者的理解和应用带来了不便；条例的篇幅过于庞大，重复的内容比较多，总则与分则之间频繁出现“参见”，某些条款中有措辞含义不清的情况，使用者难以熟练地掌握和运用；地名与人名的拼写规则采用了英美的传统方法，影响了编目条例的国际通用性，等等。但从总体上看瑕不掩瑜。

3. 中国文献著录国家标准——《文献著录总则》

《文献著录总则》（GB 3792.1—83）由我国的全国文献工作标准化技术委员会提出，全国文献工作标准化技术委员会第六分委员会起草，国家标准局 1983 年 7 月 2 日发布，1984 年 4 月 1 日起实施。在这之后，我国又颁布了与《文献著录总则》配套的适合各种类型文献著录的一系列国家标准。

1) 指导思想

我国确立的文献著录标准化的指导思想是：在著录项目的设置、著录项目的排列顺序与著录用标识符号三个方面实行“四个统一”，即中外文目录的统一、图书馆与文献情报部门目录的统一、各类型文献目录的统一、不同载体目录的统一。

我国国家标准的文献著录规则的编制，充分体现了我国制定文献工作标准的“既靠拢国际标准，又结合我国的实际”的方针政策。

也就是说，为了实现中外文目录的统一，我国的著录规则必须向国际文献著录标准靠拢。事实上，我国国家标准的文献著录规则系列就是依据 ISBD 制定的，在著录规则编制的原则及著录项目、标识符号等方面，与 ISBD 基本保持一致，其目的就是使我国与世界范围内的目录成果能够在最大程度上实现相互识别和利用的目标，为实现真正意义上的书目信息资源共享与书目控制奠定基础。

同时，为了满足我国编目工作的需要，我国国家标准的文献著录规则的编制也充分考虑

了我国的具体情况,根据我国的文献特点、读者检索习惯及文献工作部门的特定要求,适当保留了我国文献编目中的一些传统做法,并借鉴了其他国家和地区汉语言文字目录著录中一些有益的规定。

2) 编制体例

我国的文献著录标准在编制体例上采用了从总则到分则的方法,即将整套规则划分为总则和分则,并将它们分别作为单项标准制定。

总则与分则是分期陆续制定的,它们共同形成了较为完整的一套中国文献著录标准体系。具体来说,分则主要包括《普通图书著录规则》(GB 3792.2-85)、《连续出版物著录规则》(GB 3792.3-85)、《非书资料著录规则》(GB 3792.4-85)、《档案著录规则》(GB 3792.5-85)、《地图资料著录规则》(GB 3792.6-86)、《古籍著录规则》(GB 3792.7-87)等。

3) 基本结构

《文献著录总则》(GB 3792.1-83)与各个著录分则在编制结构和内容上基本保持一致,正文部分一般由引言、名词术语、著录项目、著录项目标识符和著录内容识别符、著录格式、著录详简级次、著录用文字、文献类型标识符、著录根据、著录项目细则等10项内容构成。各个著录分则还有一些附录,如标目、实例、载体名称和代码等。

4) 主要内容

主要内容包括著录项目及其标识符和著录格式两部分。

(1) 第一部分,著录项目及其标识符。具体如表4.3所示。

《文献著录总则》比较侧重传统文献的著录特征的限定。

表4.3 《文献著录总则》的著录项目及其标识符

基本描述项目	标 识 符	描述子项目
题名与责任说明项	=	正题名
	:	并列题名
	[]	副题名及说明
	/	文献类型标识
	;	第一责任者
版本项	. --	其他责任者
	/	版次及其版本形式
文献特殊细节项	.	与本版有关的责任说明
出版发行项	. --	出版地或发行地
	:	出版者或发行者
	,	出版年、月或发行年、月
	(; ,)	印刷地、印刷者、印刷年、月
载体形态项	. --	页数、卷(册)数
	:	图
	;	尺寸
	+	附件
丛编项	. -- (正丛书名
	=	并列丛书名
	:	副丛编名及说明丛编名文字

续表

基本描述项目	标 识 符	描述子项目
丛编项	/ , :)	丛编责任者 国际标准连续出版物编号（ISSN） 丛编编号
附注项		
文献标准编号及有关记载项	. -- () :	国际标准文献编号 中国标准文献编号 装订 价格
提要项		
标识项（未设置但实际使用中显示）		分类号 主题词

其中，描述项目标识符是指用以分隔描述项目的专用辅助符号。

描述项目标识符主要具有两个作用：

① 通过辅助符号的使用，可以克服语言障碍，方便地识别不同项目的含义，有利于促进不同国家地区之间的描述记录的交流和互换；

② 有利于信息组织和检索的现代化，便利机读数据处理。

（2）第二部分，著录格式。

著录格式是指著录记录内各个著录项目的记录次序和表述方式。

《文献著录总则》的著录格式如表 4.4 所示。

表 4.4 《文献著录总则》的著录格式

正题名=并列题名；副题名及说明题名文字[文献类型标识]/责任第一说明；责任其他说明.— 版本类型或版次/与本版有关的责任者 .— 文献特殊细节项. — 出版发行地：出版发行者，出版发行日期 文献数量:图表；尺寸 .— （丛编名；编次/丛编责任者）
附注项
国际标准书号（装订）：定价
提要项
排检项

如果按照以上著录格式，对《信息组织》一书进行著录，结果如表 4.5 所示。

表 4.5 《信息组织》一书著录款目样例

信息组织/马张华编著 .— 3 版. — 北京：清华大学出版社，2008.5 334 页；23cm .— （面向 21 世纪课程教材·信息管理与信息系统专业教材系列） ISBN978-7-302-17155-3:37.00 元 I. 信… II.马… III.信息管理—高等学校—教材 IV.G203

4. 《中国文献编目规则》

《中国文献编目规则》由中国文献编目规则编撰小组编辑,全国情报文献工作标准化技术委员会、中国图书馆学会推荐使用。2005年,由国家图书馆《中国文献编目规则》修订组编辑、北京图书馆出版社出版第2版。

1) 编制与修订

《中国文献编目规则》的编制,主要依据了ISBD、中国文献著录规则国家标准的原则和框架,同时参考了AACR 2和我国台湾省“中国图书馆学会”编辑的《中国编目规则(修订版)》。

《中国文献编目规则》(第2版)总的修订原则是:既遵循ISBD的原则、参照AACR 2的体例,又体现中国文献编目特色;既坚持整个编目规则体系的一致性,又考虑各种文献类型的特殊性;既坚持标准规则的统一性,又保持适当的灵活性。因此,《中国文献编目规则》修订版在坚持与国际接轨的原则的同时,还根据网络环境下电子资源的特点,尽量适应计算机编目的需要;另外,它保持了原版的结构体例,增加了新的内容,修改了不适应的条款。

《中国文献编目规则》的编制体例是将总则与各分则融于一个文献编目规则之中,一次编制完成,以章节的形式分列各类型文献著录规则的条文,在总则与各分则之间用参照的办法互相参阅,其编制体例与AACR 2基本一致。

2) 结构与内容

《中国文献编目规则》各章既互相衔接、有机联系,又自成系统。其第2版的内容结构如下:

前言

第一部分 著录法

- 第一章 总则
- 第二章 普通图书
- 第三章 学位论文、科技报告、标准文献
- 第四章 古籍
- 第五章 拓片
- 第六章 测绘制图资料
- 第七章 乐谱
- 第八章 录音资料
- 第九章 影像资料
- 第十章 静画资料
- 第十一章 连续性资源
- 第十二章 缩微文献
- 第十三章 电子资源
- 第十四章 手稿
- 第十五章 综合著录与分析著录

第二部分 标目法

- 第十六章 总则
- 第十七章 个人名称标目
- 第十八章 团体/会议名称标目
- 第十九章 题名标目
- 第二十章 参照

附录 1 著录样例

- 附录2 中国历史朝代规范简称
- 附录3 中国各少数民族规范名称表
- 附录4 世界主要国家和地区名称表
- 附录5 主要名词术语

可以看出,《中国文献编目规则》正文分为两大部分,共有20章,书末有5个附录,总计55万字。

(1) 第一部分著录法,分为15章。在这部分中,首先列出各类型文献著录原则的总则,再分别列出普通图书等13种文献的著录规则,最后一章为综合著录与分析著录规则。该部分较为全面系统地涵盖了各类型文献的著录规则,并在总体结构上为新型文献预留了篇章设置的位置。

另外,该部分每一章的内容结构也基本相同,都以通则与8个著录项目为主要内容。在通则部分,阐明使用范围、著录项目、著录用标识符号、著录详简级次、著录信息源、著录用文字等。在8个著录项目部分,详细说明各项目的具体著录规则。最后,有些章还列出一些针对特殊情况的细则。

《中国文献编目规则》的著录项目、著录标识符号与著录细则基本上都与文献著录国家标准保持一致。所不同的是,《中国文献编目规则》的修订版仍然保留了著录详简级次的内容。

(2) 第二部分标目法,分为5章。在文献描述著录的基础上,该部分为书目记录确定检索点,提供责任者名称标目与题名标目著录规则,以及标目参照编制方法,以便于编制完整的书目款目,并实现书目规范控制。

(3) 在正文的两大部分之外,《中国文献编目规则》在最后附有5种附录,对于概念或实例都有明确的说明,便于编目人员提高编目质量。

3) 主要特点

(1) 标准性。

为了便于国内外书目信息的交流,《中国文献编目规则》比较注重与国内外有关标准、规则保持协调一致。在著录法部分,全面贯彻ISBD与我国文献著录国家标准;在一些专业术语的使用方面,改变了以往传统的概念,采用了国际通行的用法。同时,修正了文献著录国家标准中与国际著录条例相违背的条款。

(2) 特色性。

由于强调了立足当代、继承传统、立足本国、借鉴国外的编制原则,《中国文献编目规则》在处理我国特有的问题时,保留了我国编目的优良传统,并不强求与国际条例一致,这个原则在标目法部分尤为突出。

(3) 丰富性。

《中国文献编目规则》在文献著录国家标准系列的基础上增加了许多新内容。例如,将国家标准《非书资料著录规则》中的“非书资料”进一步划分为录音资料、影像资料、静画资料、缩微文献、电子资源,并各自独立为一章。增加了学位论文、科技报告、标准文献、拓片、乐谱、手稿、综合著录与分析著录规则及各类型文献标目的选取与著录法等。因此,《中国文献编目规则》的内容比我国国家标准更加全面、细化,是一部完整的大型文献编目规则。

(4) 实用性。

为了便于编目人员的操作使用,《中国文献编目规则》采取了一系列措施,例如,吸取AACR2的优点,采用便于记忆的编制结构与条款编号方法;设置选择项目、弹性标目范围、交替规则等,在不违反总原则的前提下,为各类型图书馆或其他机构的编目提供较多的选择和变通的余地;增加了著录实例,并配以通俗的文字说明;设置了若干附录,提供了著录样例与规范表。

4.2 机读目录著录法

4.2.1 MARC 在全球的发展概述

1. 美国 MARC

MARC (Machine Readable Catalogue) 是“机器可读目录”,是以代码形式结构和特定结构记录在计算机存储载体上的,可用计算机识别与阅读的目录。

MARC 数据最早产生于美国。1961 年,美国国会图书馆开始图书馆自动化的设想,并于 1965 年产生了《标准机器能读目录款式的建议》(A Proposed Format for a Standardized Machine-Readable Catalog Record),1966 年 2 月开始了“MARC 试验计划”,实验结果产生了 MARC 的格式即 MARC I 格式,1967 年对 MARC I 进行改进,1968 年推出 MARC II。进入 20 世纪 70 年代,MARC 被接受为国际标准,因为是由美国国会图书馆研制的,故称 LCMARC 格式。随着 MARC II 的推出,世界许多国家和地区图书馆都相继采用 MARC II 格式建立自己的机读目录系统,为了与其他国家的 MARC 版本分开,1983 年改称为 USMARC。1994 年美国国会图书馆、加拿大国家图书馆(National Library of Canada)、大英图书馆(British Library)开始推动 USMARC、CAN/MARC、UKMARC 的 MARC Harmonization,希望整合处理英文资料的 MARC,形成一致性的机读格式^①。1999 年美国与加拿大 MARC 排除相异性联合推出 MARC 21。2001 年大英图书馆也宣布采用 MARC 21。此后瑞典、芬兰、德国、日本等国家图书馆也相继采用 MARC21 格式^②。

MARC 21 由 5 部分组成。

(1) MARC 21 书目数据格式(MARC 21 Format for Bibliographic Data):是一种用于描述、检索和控制包括图书、连续出版物(或连续性资源)、地图、电子资源、乐谱、视听资料及混合型资料在内的各类型书目资料格式。

(2) MARC 21 规范数据格式(MARC 21 Format for Authority Data):用于确认和控制那些受规范控制的书目记录中的内容和内容标识符的数据元素编码。其目的是在书目数据库中用规范文档对书目记录的检索点进行规范控制,维护书目数据库中名称和主题标目的唯一性和一致性。需要进行规范控制的检索点包括规范形式的名称、主题和主题复分^③。

(3) MARC 21 馆藏数据格式(MARC 21 Format for Holdings Data):包含各种类型资料的馆藏和位置数据。

(4) MARC 21 分类数据格式(MARC 21 Format for Classification Data):包含相关的分类号和标题等相关数据元素。分类记录可用于维护和发展分类类表^④。

(5) MARC 21 团体/社区信息格式(MARC 21 Format for Community Information):为包含事件、程序、服务等相关信息的记录提供格式规范,从而使这些信息可以被整合到公共存取目录中为用户使用。

① 张慧蓉. 浅谈 MARC 的发展趋势. <http://www.lib.ntu.edu.tw/pub/mk/mk55/mk55-09.htm>(2008-12-3).

② 编目精灵 III. 台湾采用 MARC21 格式之理由. <http://catwizard.net/posts/20120125095904.html>(2013-12-27)

③ 高红, 顾犇主编. 国家图书馆 MARC 21 格式使用手册课题组编著. MARC 21 规范数据格式使用手册[M]. 北京: 北京图书馆出版社, 2005.

④ 温琳琳. 资讯组织研究. http://203.208.37.104/translate_c?hl=zh-CN&sl=zh-TW&u=http://river.glis.ntnu.edu.tw/homework/information-1/work/MARC21_P.pdf&prev=/search%3Fq%3DMARC%2B21%2B%25E4%25B9%25A6%25E7%259B%25AE%2B%25E5%2586%2585%25E5%25AE%25B9%26start%3D20%26hl%3Dzh-CN%26newwindow%3D1%26sa%3DN&usg=ALkJrhiS3EH8n8GV-lwXEwNRDjnZKJgUiA(2008-12-2)

2. 国际标准化组织: ISO 2709

1973年ISO在审核MARC II格式后,将其格式结构定为国际标准,即《文献目录信息交换用磁带记录格式》(ISO 2709)。于1981年、1996年、2008年进行修订,改名为《信息交换格式》。它规定了书目信息机读目录的逻辑组织原则与实施方法,制定了一个通用的格式,规定了一条机读目录记录必须由头标区、目次区及数据区三大部分构成。还规定了头标区中各固定位置的含义,目次区的构成方法及标识符和分隔符的选取^①。

此格式将容纳可作目录著录的一切形式的资料的记录及权威记录这样的有关记录。本国际标准描述的是一种通用结构,是一种专门为数据处理系统之间进行交流而设计的框架,它并不要求作为系统内部的处理格式使用^②。

3. 国际图联: UNIMARC

由于各国不同的机读编目格式造成书目记录共享障碍,为了方便不同格式之间的数据交换,国际图联(IFLA, The International Federation of Library Associations and Institutions)在1977年推出UNIMARC: Universal MARC Format格式。它对图书、印刷材料及各种包括音像在内的非书资料的格式进行了规范。1980年推出第二版,这版补充制图资料所需的数据字段和更新期刊和专著的几个字段、编辑和修订其他制图属性和非书资料的临时字段。1983年印发《unimarc手册》(UNIMARC Handbook),1987年修订后发行《Unimarc手册》(UNIMARC Manual),对音像视频及缩微胶卷资料的临时数据字段进行修订并为电子资源增加临时数据字段。1994年推出第二版(UNIMARC Manual: bibliographic format),并于1996年、1998年、2000年、2002年和2005年进行5次更新。第三版已于2008年推出。除此以外,还分别于1991年推出配套的(UNIMARC/ Authorities: Universal Format for Authorities),2000年推出UNIMARC Classification Format,2003年推出UNIMARC Manual: Holdings Format^③。

UNIMARC完全符合国际标准ISO 2709的各项规定,分为头标区、目次区、数据区三个区。UNIMARC基本保持了MARC 2的结构,项目设置与内容安排大部分相同。但它力图摆脱手工著录的束缚、适应计算机处理数据的长处,进行了改进,可容纳各种类型文献,克服各国使用自己的MARC系统中的专指性并容易转换为统一的UNIMARC格式,实现共享。所以,UNIMARC作为国际机读书目数据交换格式为多国使用^④。

4. 联合国教科文组织: CCF

1978年联合国教科文组织UNESCO提出开发一种图书馆、书目机构和文摘、索引机构的通用格式,并于1984年推出公共交换格式(CCF, Common Communication Format)。

它把文献分为目标文献和相关文献,把文献的相关关系分成纵向关系和横向关系两种。被描述的主要文献称为目标文献,而与目标文献具有各种关系的所有其他文献称为相关文献。同时,它引入了区段的概念,用来在描述主要文献的同时描述多个相关文献,每个区段中存放一个文献,同一区段中有关系的字段也可以连接起来^⑤。

① 傅守灿,陈文广. 图书馆自动化基础教程[M]. 北京:北京大学出版社,1996.

② 国际标准化组织(ISO). 文献与情报工作国际标准汇编 续集一. 中国科学技术情报研究所,全国文献工作标准化技术委员会译. 北京:科学技术文献出版社,1983:7.

③ IFLA UNIMARC Core Activity PUBLICATIONS.
[http://www.ifla.org/VI/8/unimarc-publist.htm\(2008-11-16\)](http://www.ifla.org/VI/8/unimarc-publist.htm(2008-11-16)).

④ 傅守灿,陈文广. 图书馆自动化基础教程[M]. 北京:北京大学出版社,1996.

⑤ 潘岩铭 金培华. 从机读目录格式到通用记录格式:面向未来的信息服务. [http://www.marsoft.cn/marc/article6.htm\(2008-11-16\)](http://www.marsoft.cn/marc/article6.htm(2008-11-16)).

5. 中国: CNMARC

CNMARC 是中国机读目录 (China Machine-Readable Catalogue) 的缩写, 用于中国国家书目机构同其他国家书目机构及中国国内图书馆与情报部门之间, 以标准的计算机可读形式交换书目信息。中国机读目录研制开始于 20 世纪 70 年代。1979 年成立了全国信息与文献标准化技术委员会, 成立北京地区机读目录研制小组; 1982 年, 中国标准总局公布了参照 ISO 2709 制定的国家标准《文献目录信息交换用磁带格式》(GB 2901-82), 为中文 MARC 格式的标准化奠定了基础; 1986 年 UNIMARC 中译本面世。在此基础上, 根据我国实际情况, 编制了《中国机读目录通讯格式》讨论稿, 1992 年 2 月正式出版《中国机读目录通讯格式》, 即 CN-MARC。CNMARC 格式为我国机读目录实现标准化、与国际接轨, 从数据结构方面提供了保障^①。1996 年 2 月 6 日, 中华人民共和国文化行业标准《中国机读目录格式》(China MARC Format) 正式发布, 1996 年 7 月 1 日起实施。该标准根据我国文化部科技司于 1993 年 3 月向北京图书馆下达的研究任务而制定, 目的是推进书目数据的规范与统一、加速我国文献信息网络的建设和实现国内外书目信息的共建共享。2004 年, 北京图书馆出版社出版了国家图书馆编的《新版中国机读目录格式使用手册》(New China MARC Format Manual)^②。

4.2.2 MARC 记录基本格式

MARC 记录由记录结构、内容标识符和数据内容组成。记录结构遵循包括美国国家标准 Information Interchange Format (ANSI Z39.2) 和国际标准 Format for Information Exchange (ISO 2709); 内容标识符用来识别 MARC 记录的数据元素, 或提供有关数据元素附加信息的特殊符号或编码, 包括字段标识符、字段指示符和子字段代码; 数据元素内容通常遵循格式以外的其他标准, 如英美编目条例 (AACR, Anglo-American Cataloguing Rules)、美国国会图书馆主题词表 (LCSH, Library of Congress Subject Headings)^③。

下面以 MARC 21 Format for Bibliographic Data 为例说明 MARC 记录的基本格式。MARC 21 书目数据格式由记录头标区、地址目次区和可变长字段三个部分组成, 每个记录均以记录终止符结束。其结构图如图 4.1 所示。

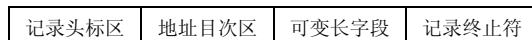


图 4.1 MARC 逻辑结构图

1. 记录头标区

记录头标区 (Leader) 位于每个记录的起始位置, 是对一条记录的总体说明, 由 24 个字符位组成固定长, 为计算机处理提供该记录的基本参数, 概括了该记录各方面的特点, 供计算机识别该记录使用。该区没有指示符和子字段代码。其 24 个字符位置分配如表 4.6 所示^{④⑤⑥⑦}。

① <http://library.gxccedu.com/wt3.html>(2008-11-16).

② 李燕, 杜薇薇, 郭华. MARC21 元数据与 CNMARC 元数据的分析比较. http://www.nlc.gov.cn/yjfw/gtcw_ywyt_bm06liy.htm(2008-12-2).

③ 高红, 顾彝. 国家图书馆 MARC21 格式使用手册课题组编著. MARC21 书目数据格式使用手册[M]. 北京: 北京图书馆出版社, 2005.

④ 刘荣. 图书情报管理自动化基础[M]. 武汉: 武汉大学出版社, 1998.

⑤ MARC21 书目格式介绍.

<http://210.32.137.28/calib/download/train/20051010/MARC21%E4%B9%A6%E7%9B%AE%E6%A0%BCE5%BC%8F%E8%AF%A6%E7%BB%86%E4%BB%8B%E7%BB%8D.doc>(2008-11-17).

⑥ MARC21 Specifications for Record Structure, Character Sets, and Exchange Media-RECORDSTRUCTURE. <http://www.loc.gov/marc/specifications/specrecstruc.html>(2008-11-17).

⑦ Leader(NR). <http://www.loc.gov/marc/bibliographic/bdleader.html>(2008-11-17).

表 4.6 记录头标区字符分配表

字符位置	名称	字符数	作用
00~04	记录总长 (Record Length)	5	记载了本逻辑记录的全部字符数, 即从记录头标区的第一个字符到本记录最后终止符的全部字符位数。代码右对齐, 不足补“0”。通常由计算机程序自动生成
05	记录状态 (Record Status)	1	反映书目记录的维护状态。共 5 种状态, 分别是: a -由简编升级的记录 (Increase in Encoding Level); c -修改过的记录 (Corrected or Revised); d -删除的记录 (Deleted); n -新记录 (New); p -由在版编目升级的记录 (Increase in Encoding Level From Prepublication)
06	记录类型 (Type of Record)	1	区别不同的资料类型和内容。有 14 种状态: a, c, d, e, f, g, l, j, k, m, o, p, r, t; a -文字资料 (Language Material); c - 乐谱 (Notated Music) (包括印刷型、缩微型和数字化乐谱); d -手稿型乐谱 (Manuscript Notated music); e -测绘制图资料 (Cartographic Material); f -手稿型制图资料 (Manuscript Cartographic Material) (如地图、图谱等); g -放映媒体 (Projected Medium) (如动画卡通、录像制品、幻灯片等); i -非音乐录音资料 (Nonmusical Sound Recording) (如演讲); j -音乐录音资料 (Musical Sound Recording) (如 CD、磁带等); k -二维非放映资料 (Two-dimensional Nonprojectable Graphic); m -电子资源 (Computer File) (包括计算机软件如程序、游戏、数值型数据、多媒体文献、联机系统或服务); o -多载体配套资料 (Kit); p - 混合型资料 (Mixed Materials); r -三维制品或天然物体 (Three-dimensional Artifact or Naturally Occurring Object); t -手稿型文字资料 (Manuscript Language Material)
07	书目级别 (Bibliographic Level)	1	用一位小写拉丁字母表示, 有 6 种状态: a, b, c, d, m, s; a -专著部分级 (Monographic Component Part); b -连续出版物部分级 (Serial Component Part); c -合集 (Collection); d -子集 (Subunit); i -集成性资源 (Integrating Resource); m -图书/单册 (Monograph/Item); s -连续出版物 (Serial)

续表

字符位置	名称	字符数	作用
08	控制类型 (Type of Control)	1	用一位字符代码表示, 有两种状态: #, a; #-无特定的控制类型 (No Specified Type), 默认为#; a-档案控制 (Archival)
09	字符编码系统 (Character coding Scheme)	1	有两种模式#, a, 默认为#; #-MARC-8; a-UCS/Unicode
10	指示符数 (Indicator Count)	1	记录中可变数据区含有多个字段, 每个字段前均有二位符号以指示该字段, 所以在此总填作 2。该字段可通过计算机程序自动生成
11	子字段代码数 (Subfield Code Length)	1	记录中可变数据区含有多个字段, 每个字段也含有若干个子字段。每个子字段前均有二位符号以指示该字段。二位符号由界定符 (\$) 和一位小写字母构成, 所以在此总填作 2。该字段可通过计算机程序自动生成
12~16	数据基地址 (Base Address of Data)	5	表示记录中第一个可变量控制字段的起始字符位置。其值等于记录第一部分记录头标区 (24 位) 和第二部分地址目次区字符数 (包括目次区的字段终止符) 总和。该字段可通过计算机程序在生成记录时自动计算产生
17	编码等级 (Encoding Level)	1	说明书目记录中的书目信息或内容标识大的完整程度。共 10 级, 分别为#, 1, 2, 3, 4, 5, 7, 8, u, z: #-完全级 (Full level), 指最完整的 MARC 记录, 编制记录时与编目实体核对过; 1-完全级 (未与编目实体核对) (Full level, material not examined); 2-次完全级 (未与编目实体核对) (Less-than-full level, material not examined); 3-简编级 (Abbreviated level); 4-核心级 (Core level); 5-部分级 (Partial (preliminary) Level); 7-最简级 (Minimal Level); 8-预编级 (Prepublication Level); u-级别不详 (Unknown); z-不应用 (Not Applicable)
18	著录标准 (Descriptive Cataloging form)	1	说明编制记录所依据的著录标准和规则, 这些标准用以下符号表示: #, a, c, I, u; #-非 ISBD 标准 (Non-ISBD); a-AACR 2; c-ISBD 标点符号省略; (ISBD Punctuation Omitted) i-包含 ISBD 标点符号; (ISBD Punctuation Included) u-不详 (Unknown)

续表

字符位置	名称	字符数	作用
19	多组成部分资源记录级别 (Multipart Resource Record Level)	1	反映了资源从属和记录之间的相互依赖关系。有以下级别:#, a, b, c: #-未指定或不适用 (Not Specified or Not Applicable); a -集合 (Set); b -有独立题名 (Part with Independent Title); c -没有独立题名 (Part with Dependent Title)
20~23 款目 布局 (Entry map)	20 字段长度的长度 (Length of Length-of-field Portion)	1	定义字段长度部分所含的字符数, 其值是 4。由计算机程序自动生成
	21 起始字符位置的长度 (Length of the Starting-character-position Portion)	1	定义起始字符位置的长度, 其值是 5。由计算机程序自动生成
	22 规定操作部分的长度 (Length of The Implementation-Defined Portion)	1	一般填 0。由计算机程序自动生成
	23 (空白)	1	未定字符, 一般填 0。由计算机程序自动生成

注: #表示空格。第 19 字符位于 2007 年对原来“连接记录要求 (Linked-record Code)”进行重新定义。“连接记录要求 (Linked-record Code)”表示在不检索相关记录的情况下, 是否能根据连接款目 (76X-78X) 生成包含相关记录的基本识别信息的附注

2. 地址目次区

地址目次区 (Directory) 紧接在记录头标区之后, 由若干个固定长度的目次区款目项及区末的字段分隔符构成。其结构如图 4.2 所示。



图 4.2 地址目次区结构

由于记录的不同, 其著录项目有多有少, 相应的字段个数的多少也就不同, 所以不同的文献书目记录, 其地址目次区的目次区款目项的个数也不完全相等。目次区的总长为: $12 \times N$ (款目项数) + 1 (字段终止符)。

地址目次区有计算机系统自动生成, 每个目次区款目项共 12 个字符位, 可分为 3 个部分: 字段标识符 (占 3 个字符位)、字段长度 (占 4 个字符位)、字段起始字符位置 (占 5 个字符位)。其具体说明见表 4.7。

表 4.7 地址目次区款目说明

字 符 位	含 义	说 明
00~02	字段标识符 (Tag)	用来识别相关字段
03~06	字段长度 (Field Length)	包含 4 位十进制数字字符, 用来识别相关字段的字符数
07~11	字段起始字符位置 (Starting Character Position)	包含 5 位十进制数字字符, 指出该字段第一个字符在记录中的字符位置, 即用来识别该字段第一个字符相对于数据基地址 (记录头标区 12~16) 所处的位置

每个款目对应记录中的一个可变长控制字段或可变长数据字段。先排可变长控制字段, 再排可变长数据字段。一个字段的起始字符位置=上个字段的起始字符位置+上个字段长度。

3. 可变长字段

可变长字段 (Variable Fields) 由两部分构成: 控制字段和数据字段。

1) 控制字段 (Control Fields)

字段标识符为 00X, 包含用于处理机读书目记录的控制号、其他控制信息和代码信息。该字段均无字段指示符和子字段代码。其基本信息如表 4.8 所示^{①②}。

表 4.8 控制字段组成表

字段标识符	名 称	必 备	可 重 复 否	备 注
001	控制号 (Control Number)	是	否	由创建、使用或发行 MARC 记录的机构设置。是 MARC 记录的唯一标识号。通常由系统自动生成
003	控制号标识 (Control Number Identifier)	是	否	包含分配 001 字段的机构代码。通常由系统自动生成
005	最近一次处理的日期和时间 (Date and Time of Latest Transaction)	是	否	由 16 位字符组成。其中日期长度 8 位字符, 格式为 yyyyymmdd。时间长度 8 位字符, 格式为 hhmmss.f 作为记录修改更新的标识
006	定长数据元素-附件特征 (Fixed-Length Data Elements-Additional Characteristics)	否	是	本字段有 18 个字符位 (00~17)。当 008 字段不能完全涵盖编目文献所有资料类型特征时, 用 006 字段作为 008 字段的补充
007	载体形态定长字段 (Physical Description Fixed Field)	否	是	采用树形结构定义, 以代码形式对非书资料或含有非书资料的文献的载体形态进行描述的定长字段 (每种资料类型定长数据元素的字符位数不同), 反映全部文献或部分文献 (如附件) 的物理特征。该字段的代码来源通常为 300 字段和 5XX 字段
008	定长数据元素 (Fixed-Length Data Elements)	是	否	以代码习惯所反映数据整体及编目文献特殊书目特征信息的定长数据字段 (40 个字符位)。主要用于数据检索和数据管理

① 高红, 顾桦. 国家图书馆 MARC21 格式使用手册课题组编著. MARC21 书目数据格式使用手册[M]. 北京: 北京图书馆出版社, 2005.

② 00X-Control Fields-General. Information. <http://www.loc.gov/marc/bibliographic/bd00x.html> (2008-12-2).

2) 数据字段

数据字段（Data Fields）包含字段指示符、子字段代码和长度不固定的子字段数据元素。其结构如图 4.3 所示。其中字段指示符提供字段内容、字段直接相互关系及数据处理中所需操作的附件信息。每个指示符的值都有特定含义，通常用数字（0~9）或空格表示。子字段代码由两个字符组成，第一个字符为界定符，用来分隔不同的子字段，第二个字符为数据元素标识符，由字母或数字组成^①。

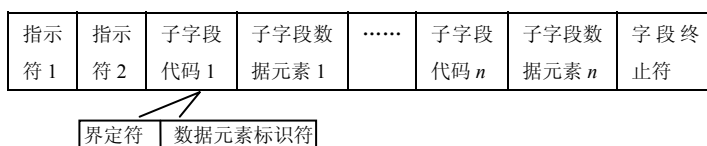


图 4.3 数据字段结构

根据可变长数据字段的第一个字段标识符可以将 MARC 21 书目数据格式分为以下字段块，具体见表 4.9^②。

表 4.9 书目记录字段块

字 段 块	含 义
0XX	控制信息、号码和代码信息（Control Information, Numbers and Codes Fields）
1XX	主要款目（Main Entry Fields）
2XX	其中 20X-24X 字段表示题名及其相关字段（Title and Title-related Fields） 25X-28X 表示版本、出版说明（Edition, Imprint, etc. Fields）
3XX	载体形态等（Physical Description, Etc. Fields）
4XX	丛编说明（Series Statements Fields）
5XX	附注（Notes Fields）
6XX	主题检索（Subject Access Fields）
7XX	70X-75X 附加款目（Added Entries Fields） 76X-78X 连接款目（Linking Entries Fields）
8XX	80X-83X 丛编附加款目（Series Added Entries Fields） 841-88X 馆藏、电子资源定位与检索等（Holdings, Location, alternate graphics etc. Fields）

注：字段的说明可以参考相关书籍及网站内容

4. 记录终止符（Record Terminator）

ASCII 的控制字符 1D（十六进制），紧跟最后一个数据字段的字段终止符后，是一个 MARC 记录的最后字符，表示该 MARC 记录的结束。

5. 样例解析

为了更好地说明 MARC 记录的构成，从武汉大学图书馆提取一条 MARC 记录，如图 4.4 所示为其磁盘数据格式，图 4.5 所示为按阅读习惯设计的字段数据加工表。

① 高红，顾桦. 国家图书馆 MARC21 格式使用手册课题组编著. MARC21 书目数据格式使用手册[M]. 北京：北京图书馆出版社，2005.

② MARC 21Format for bibliographic data.[http://www.loc.gov/marc/bibliographic/\(2013-12-2\)](http://www.loc.gov/marc/bibliographic/(2013-12-2)).

```

01756cam a2200361 a 4500
001001300000003000400013005001700017008003900034010001700073020002900090020002600119035002300145035
002000168040004400188043001200232050002500244082002400269093001400293099002100307100003300328245009
600361260006400457300003300521504006400554505031500618650004300933700002800976856008201004856009101
086856010601177950002701283999008401310-022005022343-CAL-20080630155540.0-061213s2007 enka b 001 0 eng--
a 2006101757-a9780521877657 (hardback)-a0521877652 (hardback)-a(OCOLC)ocm77116897-a(OCOLC)77116897--
aDLCcDLCdYDX-dBAKERdBTCTAdYDXCPdDLC-ad-----00aHJ5715.D44bB57 2007-00a336.2/714091724222--
aF811.424-aCAL 022008044284-1 aBird, Richard Miller,d1938--14a The VAT in developing and transitional countries /-
cRichard M. Bird, Pierre-Pascal Gendron.-aCambridge ;aNew York :-bCambridge University Press,c2007.- aix, 267 p. :-
bill. ;c24 cm.- aIncludes bibliographical references (p. 233-257) and index.-0 aWhy this book? -- The rise of VAT -- Is
VAT always the answer? -- Trade and revenue -- Equity and the informal sector -- What should be taxed? -- Key issues in
VAT design -- New issues in VAT design -- Administering VAT -- Dealing with difficulties -- The political economy of VAT
-- Where do we go from here?.- 0aValue-added taxzDeveloping countries.-1 aGendron, Pierre-Pascal.-413Table of contents
onlyuhttp://www.loc.gov/catdir/toc/ ecip076/ 2006101757.html-423Publisher descriptionuhttp://www.loc.gov/catdir/
enhancements/ fy0729/2006101757-d.html-423Contributor biographical informationuhttp://www. loc.gov/catdir/
enhancements/fy0729/2006101757-b.html-aZTF811.4/B618/2007/ Y- tEawcat2a20080630 15:55:13Gwcat2-g20080630
15:52:38Mwcat2m20080630 15:55:49-

```

图 4.4 一条 MARC 记录

```

000 01756cam a2200361 a 4500
001 022005022343
003 CAL
008061213s2007 enka b 001 0 eng
010 $a 2006101757
020 $a9780521877657 (hardback)
020 $a0521877652 (hardback)
035 $a(OCOLC)ocm77116897
035 $a(OCOLC)77116897
040 $aDLC$cDLC$dYDX$dBAKER$dBTCTA$dYDXCP$dDLC
043 $ad-----
05000$aHJ5715.D44$bB57 2007
08200$a336.2/714091724$22
093 $aF811.4$24
099 $aCAL 022008044284
1001 $aBird, Richard Miller,$d1938-
24514$aThe VAT in developing and transitional countries /$cRichard M. Bird, Pierre-Pascal Gendron.
260 $aCambridge ;$aNew York :$bCambridge University Press,$c2007.
300 $aix, 267 p. :$bill. ;$c24 cm.
504 $aIncludes bibliographical references (p. 233-257) and index.
5050 $aWhy this book? -- The rise of VAT -- Is VAT always the answer? -- Trade and revenue -- Equity and the informal
sector -- What should be taxed? -- Key issues in VAT design -- New issues in VAT design -- Administering VAT -- Dealing
with difficulties -- The political economy of VAT -- Where dowe go from here?.
650 0$aValue-added tax$zDeveloping countries.
7001 $aGendron, Pierre-Pascal.
85641$3Table of contents only$uhttp://www.loc.gov/catdir/toc/ecip076/2006101757.html
85642$3Publisher description$uhttp://www.loc.gov/catdir/enhancements/fy0729/2006101757-d.html
85642$3Contributor biographical information $uhttp://www.loc.gov/catdir/enhancements/ fy0729/2006101757-b.html
950 $aZTF811.4/B618/2007/Y
999 $tE$Awcat2$a20080630 15:55:13$Gwcat2$g20080630 15:52:38$Mwcat2$m20080630 15:55:49

```

图 4.5 便于阅读的 MARC 记录

1) 记录头标区

该记录的头标区值为：01756cam a2200361 a 4500，其含义如表 4.10 所示。

表 4.10 图 4.4 例子中记录头标区说明

字 符 位 置	字 符 数	代 码 值	说 明
00~04	5	01756	记录长度是 1756，系统自动生成
05	1	c	表示该记录被修改过，代码由编目员提供
06	1	a	表示该记录为文字资料，代码由编目员提供
07	1	m	表示书目级别为专著资料，代码由编目员提供
08	1	空格	表示该记录不包含特定的控制类型，代码由编目员提供
09	1	a	表示系统采用 UCS/Unicode 字符编码体系
10	1	2	指示符数由系统自动生成
11	1	2	由系统自动生成
12~16	5	00361	表示记录中第一个可变长控制字段的起始字符位置是 361，由系统自动生成
17	1	空格	表示该记录最完整的 MARC 记录，编制记录时与编目实体核对过。代码由编目员提供
18	1	a	表示该记录依据 AACR 2 编制，代码由编目员提供
19	1	空格	代码由编目员提供
20	1	4	由系统自动生成
21	1	5	由系统自动生成
22	1	0	由系统自动生成
23	1	0	由系统自动生成

2) 地址目次区

将图 4.4 中地址目次区内容转为表 4.11 进行说明。

表 4.11 图 4.4 例子中的地址目次区说明

款 目	字段标识符	字 段 长 度	字段起始位置
001001300000	001	0013	00000
003000400013	003	0004	00013 (=00000+0013)
005001700017	005	0017	0017 (=00013+0004)
008003900034	008	0039	0034 (=0017+0017)
010001700073	010	0017	00073 (=0034+0039)
020002900090	020	0029	00090 (=00073+00017)
020002600119	020	0026	00119 (=00090+0029)
035002300145	035	0023	00145 (=00119+0026)
035002000168	035	0020	000168(=00145+0023)
040004400188	040	0044	00188(=00168+0020)
043001200232	043	0012	00232(=00188+0044)
050002500244	050	0025	00244(=00232+0012)
082002400269	082	0024	00269(=00244+0025)
093001400293	093	0014	00293 (=00269+0024)
099002100307	099	0021	00307 (=00293+0014)
100003300328	100	0033	00328 (=00307+0021)
245009600361	245	0096	00361 (=00328+0033)
260006400457	260	0064	00457 (=00361+0096)

续表

款 目	字段标识符	字 段 长 度	字段起始位置
300003300521	300	0033	00521 (=00457+0064)
504006400554	504	0064	00554 (=00521+0033)
505031500618	505	0315	00618 (=00554+0064)
650004300933	650	0043	00933 (=00618+0315)
700002800976	700	0028	00976 (=00933+0043)
856008201004	856	0082	01004 (=00976+0028)
856009101086	856	0091	01086 (=01004+0082)
856010601177	856	0106	01177 (=01086+0091)
950002701283	950	0027	01283 (=01177+0106)
999008401310	999	0084	01310 (=01283+0027)

通过表 4.11 可知：该记录可变长字段所包含的字符数为：01310+0084=1393；其中，01310 为最后字段 999 的起始位置，0084 为该字段的长度。

该记录总长为：24（头标区长度）+（12×28+1）（地址目次区长度）+(1310+84）（可变长字段的长度）+1（记录终止符）=1756。

3）可变长字段

图 4.5 中能够很清楚地看到字段标识符及相应的内容。

① 010 \$a 2006101757。

说明：010 美国国会图书馆控制号（Library of Congress Control Number）。

定义与范围：为美国国会图书馆的书目记录号，不可重复。

指示符：未定义。

子字段：\$a—LC 控制号（LC Control Number）。

② 020 \$a9780521877657 (Hardback)。

020 \$a0521877652 (Hardback)。

说明：020 国际标准书号（International Standard Book Number）。

定义与范围：包括国际标准书号（ISBN）及其获得方式等，可重复。

指示符：未定义。

子字段：\$a - 国际标准书号（International Standard Book Number）。

③ 035 \$a(OCOLC)77116897。

说明：035 系统控制号（System Control Number）。

定义与范围：包含除 001、010、016 字段控制号外的书目机构系统控制号。

指示符：未定义。

子字段：\$a - 系统控制号（System Control Number）。

④ 040 \$aDLC\$cDLC\$dYDX\$dBAKER\$dBTCTA\$dYDXCP\$dDLC。

说明：040 编目源（Cataloging Source）。

定义与范围：包含创建原始 MARC 书目记录、分配 MARC 内容标识符、转录或修改 MARC 记录的机构代码或名称，不可重复。

指示符：未定义。

子字段：\$a-原始编目机构（Original Cataloging Agency）；

\$c-转录机构（Transcribing Agency）；

\$d-修改机构（Modifying Agency），可以重复。

⑤ 043 \$ad-----。

说明：043 地理区域代码（Geographic Area Code）。

定义与范围：也称 GAC 代码，取自 MARC Code List for Geographic Areas。更精确详尽的信息在 052 字段。包含与书目记录有关的地理区域代码，不可重复。

指示符：未定义。

子字段：\$a-地理区域代码（Geographic Area Code），查看 MARC Code List for Geographic Areas 可知 d 表示发展中国家（Developing Countries）。

⑥ 05000\$aHJ5715.D44\$bB57 2007。

说明：050 美国国会图书馆索取号（Library of Congress Call Number）。

定义与范围：由美国国会图书馆或其他机构分配给数据记录的美国国会图书馆分类体系的索取号或分类号，为可选字段，可重复。

指示符。

第 1 指示符：美国国会图书馆馆藏。

- 无信息提供；0 - 美国国会图书馆馆藏；1 - 非美国国会图书馆馆藏。

第 2 指示符：索取号来源。

0 - 美国国会图书馆分配；4 - 其他机构分配。

子字段：\$a-分类号（Classification Number）；

\$b-文献号（Item Number），由卡特号、日期、术语等组成，以区分相同分类号的不同文献。

⑦ 08200\$a336.2/714091724\$222。

说明：082 杜威十进分类号（Dewey Decimal Classification Number）。

定义与范围：提供杜威十进分类号。

指示符。

第 1 指示符：版本类型。

0 - 详版；1 - 简版；7 - \$2 字段中指定的其他版本。

第 2 指示符：分类号来源。

- 无信息提供；0 - LC 分配；4 - 其他机构分配。

子字段：\$a-分类号（Classification Number），在此为：336.2/714091724。

\$2-版本号（Edition Number），在此为 22。

⑧ 093 \$aF811.4\$24。

⑨ 099 \$aCAL 022008044284。

说明：09X 本地索书号（Local Call Numbers）。

定义与范围：被保留用于本地索书号和本地定义。

⑩ 1001 \$aBird, Richard Miller,\$d1938-。

说明：100 主要款目-个人名称（Main Entry-Personal Name）。

定义与范围：包含作为主要款目标目的个人名称，一般指个人责任者，不可重复。

指示符。

第 1 指示符：个人名称款目要素的类型（Type of Personal Name Entry Element）

0 - 名；1 - 姓；（3-家族名称。）

第 2 指示符：未定义。

子字段：\$a-个人名称（Personal name）；

\$d-与名称相关联的日期（Dates Associated with a Name），包括与名称相关的生卒年、事业鼎盛期或其他与名称相关的日期。

- ⑪ 24514\$aThe VAT in Developing and Transitional Countries /\$cRichard M. Bird, Pierre-Pascal Gendron。

说明：245 题名说明 (Title Statement)。

定义与范围：包含题目和责任说明。必备字段，不可重复。

指示符。

第1指示符：题名附加款目 (Title Added Entry)。

0 - 无题名附加款目；1 - 有题名附加款目。

第2指示符：不排档的字符 (Nonfiling Characters)。

0 - 无不排档的字符 (No Nonfiling Characters)，表示题名中没有不排档的首冠词。

1~9 不排档的字符数 (Number of Nonfiling Characters)，表示题名以不参加检索和排档的首冠词开始。其值为冠词的位数与位于题名中第一个有意义的词前面的空格、标点符号、发音符及其他特殊符号的位数之和。在此因为题名 “The VAT in Developing and Transitional Countries” 中第一个有意义的词为 “VAT”，前面有4位字符不排档 (“The”占3位+空格占1位)。

子字段：\$a-题名 (Title)；

\$c-责任说明等 (Statement of Responsibility, etc.)，根据 ISBD 著录规则，\$c 包含第一个 “/” 之后的所有数据。

- ⑫ 260 \$aCambridge ;\$aNew York :\$bCambridge University Press,\$c2007。

说明：260 出版发行项 [Publication, Distribution, etc. (Imprint)]。

定义与范围：包含与作品的出版、印刷、发行或生产等相关的信息，可重复。

指示符。

第1指示符：出版说明顺序 (Sequence of Publishing Statements)。

- 不使用/无信息提供/最早出版者；2 - 中期出版者；3 - 当前/最近出版者。

第2指示符：未定义。

子字段：\$a-出版、发行地 (Place of publication, distribution, etc.)；

\$b-出版、发行者 (Name of publisher, distributor, etc.)；

\$c-出版、发行时间 (Date of publication, distribution, etc.)。

- ⑬ 300 \$aix, 267 p. :\$bill. ;\$c24 cm。

说明：300 载体形态 (Physical Description)。

定义与范围：包括篇幅数量、尺寸、其他载体形态细节和附件信息，可重复。

指示符：未定义。

子字段：\$a-篇幅 (Extent)；

\$b-其他形态细节 (Other physical details)；

\$c-尺寸 (Dimensions)。

- ⑭ 504 \$aIncludes bibliographical references (p. 233-257) and index。

说明：504 书目等附注 (Bibliography, Etc. Note)。

定义与范围：主要记录编目文献及其附件中的书目、唱片目录、电影目录或其他参考文献，可重复。

指示符：未定义。

子字段：\$a-书目等附注 (Bibliography, etc. note)，在此表示编目文献中包含书目参考和索引，页码在括号内。

- ⑮ 5050 \$aWhy this book? -- The rise of VAT -- Is VAT always the answer? -- Trade and revenue -- Equity and the informal sector -- What should be taxed? -- Key issues in VAT design -- New issues in VAT design -- Administering VAT -- Dealing with difficulties -- The political economy of VAT -- Where do we go from here?.

说明: 505 格式化内容附注 (Formatted Contents Note)。

定义与范围: 包含一个格式化的内容附注, 通常包括不同著作的题名、文献的各部分题名, 同时也可以包括与作品或篇名相关的责任说明及卷期标识, 可重复。

指示符。

第1指示符: 显示常数控制符 (Display Constant Controller)。

0-内容附注; 1-不完整的内容附注; 2-部分内容附注; 8-不生成附注导语。

第2指示符: 内容标识级别 (Level of Content Designation)。

#-基本级; 0-增强级。

子字段: \$a-格式化内容附注 (Formatted Contents Note), 当第2指示符的值为“#”, 子字段\$a记录完整的、不完整的或部分的格式化内容。在此例中是记录完整的格式化内容。

- ⑯ 650 0\$aValue-added tax\$bDeveloping countries.

说明: 650 主题附加款目-论题性术语 (Subject Added Entry-Topical Term)。

定义与范围: 包含用作主题附加款目的论题性主题, 可重复。

指示符。

第1指示符: 主题级别 (Level of Subject)。

#-无信息提供; 0-未说明级别; 1-主要主题; 2-次要主题。

第2指示符: 主题标引叙词表 (Thesaurus)。

0-LC 主题标目; 1-LC 儿童文学主题标目; 2-MeSH; 3-美国国际农业图书馆主题规范文档; 4-未说明主题词来源; 5-加拿大主题标目; 6-加拿大法语主题标目; 7-来源在\$b指明。

子字段: \$a-论题性术语或地理名称款目要素 (Topical Term or Geographic Name Entry Element);

\$b-地理复分 (Geographic Subdivision)。

- ⑰ 7001 \$aGendron, Pierre-Pascal.

说明: 700 附加款目-个人名称 (Added Entry-Personal Name)。

定义与范围: 包含作为附加款目标目的个人名称, 一般指不适合著录在 600 字段或 800 字段的分担责任者、混合责任者等, 可重复。

指示符。

第1指示符: 个人名称款目要素的类型 (Type of Personal Name Entry Element)。

0-名; 1-姓; 3-家族名称。

第2指示符: 附加款目类型 (Type of Added Entry)。

#-无信息提供; 2-分析款目。

子字段: \$a-个人名称 (Personal Name)。

- ⑱ 85641\$3Table of contents only\$uhttp://www.loc.gov/catdir/toc/ecip076/2006101757.html.

- ⑲ 85642\$3Publisher description\$uhttp://www.loc.gov/catdir/enhancements/fy0729/200610175_7-d.html.

- ⑳ 85642\$3Contributor biographical information \$uhttp://www.loc.gov/catdir/enhancements/fy0729/2006101757-b.html.

说明: 856 电子资源定位与访问 (Electronic Location and Access)。

定义与范围: 包含定位与访问电子资源所需的信息。既可用于电子资源的定位与访问, 也可用于编目文献的电子版或相关电子资源的定位与访问, 可重复。

指示符。

第1指示符: 访问方法 (Access Method)。

#- 无信息提供; 0- 电子邮件; 1- 文献传输协议; 2- 远程登录; 3- 拨号上网; 4- 超文本协议; 7- 访问方法在\$2 说明。

第2指示符: 关系 (Relationship)。

#- 无信息提供; 0- 资源;

1- 资源版本, 表示编目文献不是电子资源, 本字段所定位的是编目文献的电子版;

2- 相关资源, 表示编目文献不是电子资源, 本字段所定位的是与该文献相关的电子资源;

8- 不生成附注导语。

子字段: \$3- 专指资料 (Materials Specified), 用于标识本字段所应用的编目文献中某一特定部分信息。当第2指示符为1时, 由\$3指明相关的部分; 当第2指示符为2时, 由\$3指明电子资源与编目文献整体的相互关系。

\$u- 统一资源标识符 (Uniform Resource Identifier)。

②1 950 \$aZT\$f811.4/B618/2007/Y。

说明: 950 是武汉大学图书馆自定义的索书号字段。

②2 999 \$tE\$Awcat2\$a20080630 15:55:13\$Gwcat2\$g20080630 15:52:38\$Mwcat2\$m 20080630 15:55:49。

说明: 999 也是武汉大学图书馆自定义字段, 记录一条书目记录从创建、修改到审校的工作轨迹。

4.2.3 MARC 著录的优缺点

1. MARC 的优点

MARC 格式是在传统的文献编目工作中发展起来的, 它的诞生规范了图书馆书目信息, 推动了图书馆自动化的发展。它的优点简述如下。

(1) 著录信息丰富。MARC 格式提供标题、责任声明、版次、出版信息、载体形态、主题词、标准代码等描述信息, 并设置主要款目字段 (1XX)、丛编说明字段 (4XX)、主题检索字段 (6XX)、其他附加款目字段 (7XX) 及丛编附加款目字段 (8XX) 等检索点信息。除了书目信息外, MARC 格式还支持视听资料、计算机软件、图像等非图书资源及网络资源的著录, 如 856 字段“电子资源定位与访问”, 可以对网络资源进行著录。

(2) 实现规范控制。MARC 格式中 LCSH 和 DDC 等国际标准和准则进行著录标目的规范控制, 提高目录编制和检索的效率、确保目录质量、完善目录功能。

(3) 方便访问, 便于计算机处理。如果一条书目记录被正确标注并储存在计算机中, 通过编写计算机程序不仅可以灵活显示或打印信息, 还可以提供对特定字段的多途径检索, 各个图书馆可以根据自己的需要确定自己系统的检索点。

(4) 便于实现资源共享。编目信息共享是一个现实问题。许多图书馆员希望获取而不是创建这些编目信息。MARC 格式已成为图书馆情报行业内的通行标准, 现行的图书馆系统大多支持 MARC 格式, 从而可以有效地实现数据共享、联合编目和联合目录的应用, 使各个

图书馆的数据在世界范围内“可见”。同时在编目信息共享的情况下,可以改变各图书馆编目详简级次不同、水平高低不一的情况,提高各个图书馆编目的质量。

(5) 有专门机构进行维护和修订。随着图书馆自动化技术和编目需求的变化, MARC 格式和相关文档也需要相应的改变。虽然 MARC 格式是由美国国会图书馆推出并发行的,但并不是单独由它进行修订的。从 1973 年开始,美国图书馆协会(ALA)下的一个跨部门委员会 MARBI(机读书目信息委员会, Machine-Readable Bibliographic Information Committee)和 MARC 顾问委员会(MARC Advisory Committee)负责审查和修订 MARC 21 文档。每次 ALA 会议都会就 MARC 格式的变化或发展所提交的讨论稿和建议进行审查。2013 年 MARBI 解散后, MARC21 的相关活动完全由 MARC 顾问委员会负责^①。

(6) 提高图书馆工作效率。MARC 格式的使用,不仅有利于编目工作的规范化、标准化,还可以从多方面提高图书馆的效率。例如,对人员配置来说,统一的编目体系及共享方式,可以使一些小型的或不具备高资料编目人员的图书馆直接使用他馆的共享编目资源;现有的 MARC 编目中心大多提供书刊,因而各图书馆可以采用“藏书来源控制法”实施图书采购的控制,优化藏书结构,避免书刊经费的浪费^②。

2. MARC 的缺点

随着计算机和网络技术的持续发展,图书馆书目数据和自动化系统质量问题也凸显出来。前者主要集中在 MARC 数据的质量、MARC 字段的适用性等方面;后者不仅受前者的影响,系统本身的设计质量、功能问题也受到考验,如系统是否装载完整记录、内容标识符是否能显示、是否允许记录下载或重新写回数据文件等。

网络资源的剧增、新型媒介的出现及电子出版的迅速发展对现有的编目规则和 MARC 格式提出了新的挑战。2002 年 Roy Tennant 曾在 Library Journal 上发文号召废止 MARC,引发了 MARC 是废除还是保留的一场大讨论^③。总结起来, MARC 格式有以下缺点。

(1) 编目人员的高要求性。MARC 格式的最大特点就是能充分揭示文献特征,提供多途径的检索,其标准化、规范化程度都对编目人员提出了较高要求。编目员不仅要熟悉 MARC 机读目录著录格式,还要掌握分类法、主题词表的发展情况,同时要在掌握标引规则的基础上注重分类和主题标引的一致性^④。对非专业人员来说, MARC 格式过于复杂。这无疑限制了编目人员的范围^⑤。

(2) MARC 格式著录单元的限制。MARC 格式的各个数据单元专指性强,字段众多,定义严格,造成 MARC 格式在某种程度上的烦琐与复杂。存在着字段之间内容的重复性、条目之间的关联性及对网络资源的适用性等问题。如 100 字段\$a 子字段用来设置个人名称的款目, 245 字段\$c 子字段同样反映了该信息,存在内容的重复; 856 字段可以实现书目与网络资源的连接,但是由于网络资源的动态性,导致网络资源地址的变更,而这种变更若不能及时反映,则会导致数据的不可获得性,并增加对书目数据库的维护难度。

(3) MARC 的标识系统的难阅读性。MARC 格式中的字段、子字段都采用了代码进行标识。这种标识系统不够直观,一般用户是难以阅读理解的,也不能使用通用的搜索引擎进行有效的检索^⑥。

① MARC Advisory Committee. <http://www.loc.gov/marc/mac/advisory.html>(2013-12-23)

② 黄建年. MARC 数据与图书馆[J]. 津图学刊, 1997(4):26~32.

③ 吴万晔. 论 MARC 元数据的缺陷及发展趋势[J]. 图书馆工作与研究, 2006(2):28~29.

④ 吴翠兰. 当前图书馆编目工作的发展契机. <http://www.chnlib.com/Zy1wj/2709.html>(2008-12-4).

⑤ 孙更新. 文献信息编目[M]. 武汉: 武汉大学出版社, 2006:454.

⑥ 延卫平. MARC 的 XML 交换格式研究[J]. 现代图书情报技术, 2006(8):31~35.

(4) 对多媒体信息描述不够。图像、音频、视频等多媒体信息已成为互联网的主体, MARC 格式虽然能够对这些资源进行描述, 但是深度不够, 不能满足多媒体信息存取需求, 对于数字图书馆数据资源的整合构成障碍^①。

(5) 使用环境和范围的限制。MARC 是图书馆系统的专用格式, 只有在符合 MARC 格式的应用前端(如图书馆专用的 OPAC)和支持 MARC 格式的专用的 Z39.50 协议才能准确获取 MARC 数据, 才能创建和使用 MARC 记录, 导致无法通过 Internet 浏览和检索目前大量存在的 MARC 格式的书目数据, 从而严重制约了这些数据的利用^②。

(6) FRBR(书目记录的功能需求)对 MARC 的冲击。国际图联于 1998 年正式推出 FRBR(FRBR, Functional Requirements for Bibliographic Records)报告, 是国际编目原则和编目思维模式上的重大突破, FRBR 认为编目对象不能停留在传统的平面层次上, 应根据用户的需求将编目对象分成若干层次, 它揭示了隐匿在编目对象中的深层次关系, 形成一个立体的元数据模型, 已经成为人们设计、考查和评估元数据的一个研究框架。虽然 MARC 本身的结构复杂, 但是 MARC 是以卡片目录为基础而发展来的, 脱胎于卡片目录, MARC 的一大缺陷是受制于卡片目录的思维, 拘泥于卡片目录显示的格式, 对复杂的层级关系的适应性却较弱, 难以适应 FRBR 的多层次等级立体新框架^③。

4.3 元数据著录法

4.3.1 元数据简介

1. 元数据的定义

元数据(Metadata), 又叫做“描述数据”或“诠释数据”。简单来说, 元数据就是“关于数据的数据”。元数据并不是个陌生的概念。我们在日常的工作中有意或无意地使用了元数据。例如, 地图的图例、图书目录等。地图的图例描述了地图中的符号的含义, 图书目录描述了图书的主要内容、图书的作者及图书存在于图书馆中的位置等。从一般的意义上讲, 因为它们就是对数据的描述, 所以这些都是元数据。同时我们也可以看出, 元数据的内容是十分广泛的。元数据的内容要根据实际的需要来确定^④。

在不同领域, 对元数据的定义也是不同的。例如, 在数据仓库领域中, 元数据被定义为: 描述数据及其环境的数据。在软件构造领域, 元数据被定义为: 在程序中不是被加工的对象, 而是通过其值的改变来改变程序的行为的数据。在图书馆与信息界, 元数据被定义为: 提供关于信息资源或数据的一种结构化的数据, 是对信息资源的结构化的描述^⑤。其中有代表性的定义有如下几种^⑥。

(1) 数据的数据(Data About Data)。

(2) 结构化数据(Structured Data about Data)。

(3) 元数据由描述数据特征的信息构成(Metadata Consist of Information that Characterizes

① 韩立栋. 基于 XML 的 MARC 探讨[J]. 兰台世界, 2005(13):60~61.

② 郦金花. 基于 CNMARC 字典库和 XML 的 MARC 发布系统的设计[J]. 情报杂志, 2006(12):87~89.

③ 吴万晔. 论 MARC 元数据的缺陷及发展趋势[J]. 图书馆工作与研究, 2006(2):28~29.

④ 中国 21 世纪议程管理中心. Internet 与可持续发展网络实用教程[M]. 北京: 科学出版社, 1998.

⑤ 元数据. <http://baike.baidu.com/view/107838.htm>(2008-12-7).

⑥ 刘炜. 关于元数据的十万个为什么. <http://www.libnet.sh.cn/sztsg/fulltext/abc/metaFAQ.pdf>(2008-12-7).

Data)^①。主要包含描述数据的内容(What)、覆盖范围(Where, When)、质量、管理方式、数据的所有者(Who)、数据的提供方式(How)等信息,是数据与数据用户之间的桥梁。

(4) 资源的信息(Information about a Resource)。

(5) 编目信息(Cataloguing Information)。

(6) 管理、控制信息(Administrative Information)。

(7) 是一组独立的关于资源的说明(Metadata Is a Set of Independent Assertions about a Resource)。

(8) 定义描述其他数据的数据(Data That Defines and Describes Other Data)(ISO/IEC 11179-3:2003(E))。

(9) 关于数据的内容、质量、环境和其他特征的数据(Data About The Content, Quality, Condition, and Other Characteristics of Data)(FGDC 标准)。

(10) 元数据是在应用中或某种环境下提供关于其他数据的信息或说明的定义数据(In data processing, metadata is definitional data that provides information about or documentation of other data managed within an application or environment)^②。

本书对元数据定义是:元数据是按照一定的标准,规范化描述一个具体的资源对象的各项特征元素集,通过这组特征元素集实现能对这个资源对象的定位、发现与获取等功能。

2. 元数据的功能

(1) 描述。这是元数据的最基本职能。是指通过对信息资源的描述,揭示信息资源的形式特征和内容特征。描述的详细与深入程度则根据不同元数据格式而不同。

(2) 定位。主要是指通过对信息资源的位置信息的描述,帮助人们了解信息资源所在之处的信息,方便信息资源的获取。此外,一旦确定信息资源的位置元数据后,也可以确定该信息资源在整个信息资源集合中的位置,这是定位的另一层含义。

(3) 检索。在著录过程中,通过描述信息资源的主要特征,并赋予检索点,建立信息资源之间的联系,有利于从多途径、多角度检索到该信息资源。

(4) 选择。通过对信息资源的描述,使用户无须浏览信息资源本身,就能对信息资源的内容有所了解、认识,结合使用环境,用户可以选择符合要求的资源。

(5) 评估。利用统计工具,对信息资源的版本、使用、保存管理等信息进行统计分析,方便资源的建立与管理者更好地组织资源,了解该信息资源在同类资源中的重要性。

(6) 管理。元数据元素除包含比较全面的著录描述信息外,还往往包括权利管理、电子签名、资源评鉴、使用管理、支付审计等管理方面的信息。

(7) 保存。元数据中往往包括详细的格式信息、制作信息、保护条件、转换方式、保存责任等内容,从而支持对资源的保护与长期保存。

3. 元数据的类型

从资料来看,元数据的划分主要有以下几种方法。

1) 按功能分

按功能分也有几种,一种是分为3个较宽泛的类:描述型元数据、结构型元数据和管理型元数据^③。一种是分为4个类:描述型元数据、结构型元数据、管理型元数据和保存型元

^① Frequently-asked questions on FGDC metadata. [http://geology.usgs.gov/tools/metadata/tools/doc/faq.html#q1.1\(2008-12-7\)](http://geology.usgs.gov/tools/metadata/tools/doc/faq.html#q1.1(2008-12-7)).

34 Metadata. [http://en.wikipedia.org/wiki/Metadata\(2008-12-7\)](http://en.wikipedia.org/wiki/Metadata(2008-12-7)).

^② [美]Arlene G. Taylor. 信息组织[M]. 张素芳,等,译. 北京:机械工业出版社,2006.

数据。还有一种也是分为4类,分别是:内容元数据、管理元数据、负载元数据和参考元数据^①。还有的是分为5个类:描述型元数据、管理型元数据、保存型元数据、技术型元数据和使用型元数据^②。在此采用较宽泛的3类,其他类可以归结为子类。

(1) 描述型元数据。

描述型元数据是指那些用来描述和识别信息资源的特征的元数据。包括题名、作者、出版日期、主题词、分类号及资源之间关系等元数据。

(2) 管理型元数据。

管理型元数据是指用来维护和管理信息资源的元数据。包括信息资源的版本信息、使用权限、获取条件和方式、数字签名、历史保存信息等。

(3) 结构型元数据。

结构型元数据是指那些文件的结构或“标记”、数据集和其他被描述的信息体。用于确保数字化信息体正常发挥功能的技术性信息。结构型元数据有时也被称为技术型元数据、显示型元数据或使用型元数据。它是指那些相关文件如何组成在一起和对对象如何在各种系统间显示和发布的相关信息。它所处理的是研究对象是什么、该对象能做什么、该对象的工作原理是什么等相关的信息。包括软/硬件文档、技术型信息等^③。

2) 按元数据格式的结构复杂程度分

登普西和希里将元数据的格式分成三组。

(1) 全文索引,是来自资源本身的全文索引的数据,通常情况下,不使用“无元数据”这一术语。

(2) 简单结构化格式,如都柏林核心元数据集等。

(3) 那些结构较为复杂或专业性较强的格式,如FGDC、GILS等。

4.3.2 都柏林核心元数据集

1995年3月,由OCLC与美国国家超级计算应用中心(NCSA)联合主持,52位来自图书馆界、电脑网络界专家共同研究产生了都柏林核心元素集(DC, Dublin Core Element Set)。旨在通过建立一套描述网络资源的元素集合,来支持网络检索^④。DC的维护和持续发展由都柏林核心倡议(DCMI, Dublin Core Metadata Initiative)组织负责,其网站为<http://dublincore.org>^⑤。

1. DC元数据的基本原则

在第一届DC会议中,提出DC元数据发展和设计的基本原则。其具体内容如下所述^{⑥⑦⑧}。

(1) “简单易用性原则”(Simplicity of Creation and Maintenance)。DC要求定义一个能得到最广泛应用、被全球所理解和接受的最小元素集,并能作为特殊用户详细描述需求的一

① 宋炜,张铭. 语义网简明教程[M]. 北京: 高等教育出版社, 2004.

② 刘嘉. 网络信息资源的组织——从信息组织到知识. [M]. 北京: 组织北京图书馆出版社, 2002.

③ [美]Arlene G. Taylor. 信息组织[M]. 张素芳,等,译. 北京: 机械工业出版社, 2006.

④ 刘炜,楼向英,赵亮. DC元数据的历史、现状及未来. [http://eprints.rclis.org/archive/00003408/01/DCMI4%E5%B9%B4%E5%88%8A_DC.pdf\(2008-11-17\)](http://eprints.rclis.org/archive/00003408/01/DCMI4%E5%B9%B4%E5%88%8A_DC.pdf(2008-11-17)).

⑤ 孙更新. 文献信息编目[M]. 武汉: 武汉大学出版社, 2006.

⑥ 刘炜,楼向英,赵亮. DC元数据的历史、现状及未来. [http://eprints.rclis.org/archive/00003408/01/DCMI4%E5%B9%B4%E5%88%8A_DC.pdf\(2008-11-17\)](http://eprints.rclis.org/archive/00003408/01/DCMI4%E5%B9%B4%E5%88%8A_DC.pdf(2008-11-17)).

⑦ 段明莲,沈正华. 数字时代的图书馆信息资源组织[M]. 北京: 北京图书馆出版社, 2006.

⑧ Using Dublin Core. [http://dublincore.org/documents/usageguide/index.shtml\(2008-12-25\)](http://dublincore.org/documents/usageguide/index.shtml(2008-12-25)).

个核心集。同时 DC 要求能方便作者和信息提供者描述自己的文档，而不给他们增加太多的负担，并能方便地实现资源发现工具之间的互操作性。

(2) “内在性原则” (Intrinsicality)。指 DC 元数据以揭示描述对象自身的内容属性为主、外部属性为辅。所谓揭示内在本质的数据，是指描述信息资源能够发现的知识内容和物理形式。

(3) “可扩展性原则” (Extensibility)。希望 DC 成为一个“核心”元素集合且可以通过各种方式扩展为适应各领域资源描述需要的元数据方案。有利于用户为特殊的目的或要求，额外增加一些著录元素，扩展元数据的著录信息。这一扩展机制，有利于 DC 元数据集随着时间的推移不断地得以修订和完善。与此同时，确保与原来定义的元素集向后的兼容性。

(4) “句法独立原则” (Syntax Independence)。指 DC 元数据的元素可以以多种方式编码应用于各类技术平台中。DC 只规定元素的基本语义。

(5) “可选择性” (Optionality)。指 DC 元素集中的任何元素都是可选的，从而简化著录元素。

(6) “可重复性” (Repeatability)。指 DC 元素集中的任何元素都是可重复的。解决了多题名、多著者、多主题等信息资源的著录问题。

(7) “可修改性” (Modifiability) 和“向上兼容原则” (Dumb-Down Principle)。这两条原则都是针对修饰词的。前者指在具体应用中可以对 DC 元素集中的任何元素进行进一步“修饰”或“限定”，但不能扩大或改变元素的基本语义。后者主要指元素的修饰词在使用中可以被忽略，而元素的值仍可作为术语被使用，即修饰词的语义包含于未修饰词中。例如，对于 Date 元素，可以忽略它的修饰词，其值仍然可以作为一个 Date 值使用，只是精确性可能会有影响。通过对元素语义的限定和修改，以便满足不同行业领域的需求。

(8) “1:1 原则” (One-to-One Principle)。一般而言，DC 元数据描述的是一个资源的一种表现形式或版本，而不是各种表现形式的相互替代。例如一幅油画，它有一个数字版本是 JPEG 的图像，在著录时，对于数字图像就应该单独著录，而且除了提供油画的作者信息外还要提供该数字图像制作者信息。简而言之，该原则要求一条描述中的每个属性必须是所描述资源的一个特性，一条元数据描述仅描述一个资源。

(9) “合适取值原则” (Appropriate Values)。特定元素或修饰词的使用应该随使用环境而改变。但是使用者不能预言出由机器翻译解释的元数据。因此，为了确定语义，必须在元数据构建时给出约束条件。

(10) “国际性原则” (International Scope)。主要是指 DC 元数据的发展要考虑到电子信息资源的多语种和多文化特征。

2. DC 的元素构成

都柏林核心标准包括简单 (Simple) 和限定 (Qualified) 两个层次。简单的都柏林核心包括 15 个元素，这 15 个元素都是可选的和可重复的，其定义如表 4.12 所示^①。限定的都柏林核心包括 3 个额外的元素 (Audience、Provenance 和 RightsHolder)，以及一组元素的限定词 (Qualifiers，或称 Refinements)，用于在资源发现上限定元素的语义要素。

表 4.12 简单 DC 元素表

元 素 名	中文名称	定 义	备 注
Title	资源名	赋予资源的名称	一般而言，这一名称指的是资源对象的正式公开的名称

① 都柏林核心元数据元素集，1.1 版本。http://dc.library.sh.cn/dces1999.htm (2008-12-22)。

续表

元 素 名	中 文 名 称	定 义	备 注
Creator	创建者	创建资源内容的主要责任者	创建者的实例包括个人、组织或某项服务系统。一般而言,用创建者的名称来标识这一条目
Subject	主题和关键词	有关资源内容的主题描述	如果要描述特定资源的某一主题,一般而言,采用关键词、关键词短语或分类号
Description	说明	对资源内容的说明	说明元素可以包括但不限于以下部分:摘要、目录、对以图形揭示内容的资源而言的文字说明或者一个有关资源内容的自由文本描述
Publisher	出版者	使资源成为可获得状态的责任者	出版者的实例包括个人、组织或某项服务系统。一般而言,用出版者的名称来标识这一条目
Contributor	其他责任者	对资源内容创建做出贡献的其他责任者	其他责任者的实例包括个人、组织或某项服务系统。一般而言,用其他责任者的名称来标识这一条目
Date	日期	与资源本身生命周期中的一个事件相关的日期	一般而言,日期应与资源的创建或可获得的日期相关。建议采用的日期格式应符合 ISO 8601[W3CDTF]规范,并使用 YYYY-MM-DD 的格式
Type	资源类型	有关资源内容的特征或类型	资源类型包括描述资源内容的一般范畴、功能、流派或聚类层次的术语。建议采用来自于受控词表中的值(如都柏林核心的资源类型指导性草案[DCT1])。要描述资源的物理或数字表现形式,请使用格式(FORMAT)元素
Format	格式	资源的物理或数字表现形式	一般而言,格式可以包括资源的媒体类型或大小。格式元素可以用来决定显示或操作资源所需的软件、硬件及其他的相应设备。如大小包括资源所占的存储空间及持续时间。建议采用来自于受控词表中的值,(如互联网媒体类型表[MIME]定义了计算机媒体格式)
Identifier	资源标识符	在特定范围内给予资源的一个明确的标识	建议对资源的标识采用符合某一正式标识体系要求的字符串或数字。例如,统一资源标识符(URI)、资源定位符(URL)、数字对象标识符(DOI)和国际标准书号(ISBN)都是正式的标识体系
Source	来源	对一个资源的参照,当前资源源自这一参照资源	当前资源可能部分或全部源自来源元素所标识的资源。建议对这一资源的标识采用一个符合正式标识体系的字符串或数字
Language	语种	描述资源知识内容所使用的语种	建议本元素的值采用 RFC 4646 标准中所定义的语种代码的规范
Relation	关联	对相关资源的参照	建议对关联的标识采用符合正式标识体系的字符串或数字
Coverage	覆盖范围	资源内容所涉及的范围	典型的范围包括空间位置(一个地名或地理坐标)、时间段(一个时间标签,日期或一个日期范围)、管辖范围(如已命名的行政实体)。范围元素的值最好取自一个受控词表(如地理名称词表[TGN]),并应尽可能地使用由数字表示的坐标或日期区间来描述地名与时间段
Rights	权限	有关资源本身所有的或被赋予的权限信息	一般而言,权限元素应包括一个对资源的权限管理声明,或者是对可提供这一信息的服务机构的参照。一般包括知识产权(IPR)、版权或其他各种各样的产权。如果没有权限元素的标注,不应与此资源相关的上述权利或其他权利做出任何假定

3. DC 的修饰词

DC 有两类：一类是元素限定词（Element Refinements）也叫修饰词，是由特定的词汇对元数据元素语义的进一步限定和细化，使其具有专指性；另一类是编码体系限定词（Encoding Scheme），它有助于对元素修饰词值的理解，这类体系包括控制词表及正规的符号或解读方式。值的表示采用来自控制词表的表记符号（如分类体系或主题词表的术语）或者具有特定含义及一定形式组成的字符串。一种编码体系无法被客户机或代理所理解，它的值仍能被其他人所理解。用于修饰的编码体系必须是清晰明确的说明，并能为公众所获取^①。DC 的主要修饰词及其含义如表 4.13 所示^{②③}。

表 4.13 DC 主要限定词

元 素 名	限 定 词	含 义	编 码 方 案
Title	Alternative	任何可替代正式题名的其他名称	—
Creator	—	—	—
Subject	—	—	LCSH MeSH DDC LCC UDC
Description	abstract	资源内容的概要	—
Description	Table Of Contents	资源内容的子单元列表	—
Publisher	—	—	—
Contributor	—	—	—
Date	Available	可获得日期。资源将在这段时间内可以获得或曾经可以获得（通常是一个时间区间）	DCMI Period W3C-DTF
Date	Created	资源创建的日期	
Date	Date Accepted	接收资源的日期（如大学院系收到的论文，期刊收到的文章等）	
Date	Date Copyrighted	版权声明的日期	
Date	Date Submitted	资源（文章和论文等）的递交日期	
Date	Issued	资源正式发布（如出版）的日期	
Date	Modified	资源被修改的日期	
Date	Valid	资源生效日期（通常是一个时间区间）	
Format	Extent	资源的大小或持续时间	—
Format	Medium	资源的物质载体或组成材料	—
Identifier	Bibliographic Citation	资源生效日期（通常是一个时间区间）	—
Identifier	—	—	URI
Source	—	—	URI
Language	—	—	ISO 639-2RFC 3066

① 陈耀盛. 网络信息组织[M]. 北京: 科学技术文献出版社, 2004.

② DCMI 元数据术语. <http://dc.library.sh.cn/dcmi-terms.htm#part3> (2008-12-22).

③ <http://dublincore.org/documents/usageguide/qualifiers.shtml> (2008-12-24).

续表

元 素 名	限 定 词	含 义	编 码 方 案
Relation	Conforms To	对资源所遵循的已有标准的参照	URI
Relation	Has Format	格式转换。所描述的资源在被参照的资源之前出现, 参照资源在实质上与所描述资源有着相同的知识内容, 只是格式不同	
Relation	Has Part	所描述的资源在物理或逻辑上包含被参照的资源	
Relation	Has Version	版本关联。所描述的资源有译本、修改本或改编本等, 也就是被参照的资源	
Relation	Instructional Method	指导方法。用于生成知识、意见和技巧的处理方法, 资源用来支持此处理方法	
Relation	Is Format Of	所描述的资源与被参照的资源有相同的知识内容, 但用另一种格式表现出来	
Relation	Is Part Of	所描述的资源是被参照资源物理或逻辑上的一个组成部分	URI
Relation	Is Referenced By	被参照的资源参考、引用或以另外的方式指引所描述的资源	
Relation	Is Replaced By	所描述的资源已被参照的资源所代替、替换或取代	
Relation	Is Required By	所描述的资源对于被参照资源而言或者在逻辑上、或者在物理上是必不可少的	
Relation	Is Version Of	所描述的资源是被参照资源的译本、修订本或改编本。版本的变化意味着是内容而不是格式有了实质的改变	
Relation	References	所描述的资源参考、引用或以其他方式指引了被参照资源	
Relation	Replaces	所描述的资源代替、替换或取代了被参照的资源	
Relation	Requires	所描述的资源需要被参照资源支持其功能、传递或在内容上保持一致	
Coverage	Spatial	所描述资源知识内容的空间特征	DCMI Point ISO 3166 DCMI Box TGN
Coverage	Temporal	所描述资源知识内容的时间特征	DCMI Period W3C-DTF
Rights	Access Rights	关于谁能访问资源的信息, 或者是对资源密级状态的说明	—
Rights	License	允许对资源进行操作的官方法律文件	URI

4. DC 的创建

1) HTML 编码

(1) META 标签。

META 标签可以对一个有名称的元数据元素进行编码。其基本语法是:

```
< meta name = "PREFIX.ELEMENT_NAME" content = "ELEMENT_VALUE">
```

例如: <meta name = "DC. Creator " content = "Taylor">

其语法含义是: PREFIX 为 DC, ELEMENT_NAME 为 Creator, ELEMENT_VALUE 为 Taylor。其含义为元素 Creator 是在 DC 元素集中定义的, 而 Taylor 是 Creator 的具体值。

对于 DC 修饰词，一般有以下三种方式。

① `<meta name="DC.ELEMENT_NAME">`

`lang="LANGUAGE_OF_METADATA_CONTENT"`

`content="ELEMENT_VALUE">`

表示该 DC 元素描述所使用的语言。

例如：`<meta name="DC.Title"`

`lang="en"`

`content="The Green Table and the Red Chair">`

表示 Title 元素使用的是英语（en）描述，其内容为“The Green Table and the Red Chair”。

② `<meta name="DC.ELEMENT_NAME"`

`scheme="CONTROLLED_FORMAT_OR_VOCABULARY_OF_METADATA"`

`content="ELEMENT_VALUE" >`

表示该 DC 元素值的编码体系。

例如：`<meta name="DC.Language"`

`scheme="rfc1766"`

`content="es">`

表示采用 RFC 1766 中定义的语种代码为 es（西班牙语）。

③ `<meta name="PREFIX.ELEMENT_NAME.SUBELEMENT_NAME" ... >`

是对有修饰词的元素的描述。

例如：`<meta name="DC.Date.Created content="1935">`

表示创建日期是 1935 年。

（2）LINK 标签。

LINK 标签用于建立一个与其他文档的联系，可以用 LINK 标签前缀词（Prefix）与它的相关定义做出参照^①。如果没有 LINK 标签与相应的定义文档关联，只有 META 标签描述的资源是不完整的^②。其基本语法是：

`<link rel="schema. PREFIX" href="LOCATION_OF_DEFINITION">`

例如：`<link rel="schema.DC" href="http://purl.org/DC/elements/1.0/">`，其语法表示是：PREFIX 为实际使用的前缀，在此值为 DC，LOCATION_OF_DEFINITION 为定义文档的 URL 或 URN，在此值为 `http://purl.org/DC/elements/1.0/`。说明所使用的 DC 的 URL 为 `http://purl.org/DC/elements/1.0/`。

2) XML 编码

在使用 XML 对 DC 元素进行描述的时候，DCMI 有以下推荐^③：

① 使用 XML Schemas 进行描述。XML Schemas 提供了包括元数据在内的 XML 文档结构定义方法。

② 使用命名空间唯一确定 DC 的元素、修饰词和编码体系。命名空间及其 Schema 地址如表 4.14 所示^④。

① 吴建中. DC 元数据[M]. 上海：上海科学技术文献出版社，2000.

② 在 HTML 中使用 DC 元数据. <http://www.chinaipower.com/A200508/2005-08-02/178448.html> (2008-12-29).

③ Guidelines for implementing Dublin Core in XML. <http://dublincore.org/documents/2003/04/02/dc-xml-guidelines/> (2008-12-29).

④ <http://dublincore.org/schemas/xmls/> (2008-12-29).

表 4.14 命名空间和地址

目标命名空间	Schema 地址
http://purl.org/dc/elements/1.1/	http://dublincore.org/schemas/xmls/qdc/dc.xsd
http://purl.org/dc/terms/	http://dublincore.org/schemas/xmls/qdc/dcterms.xsd
http://purl.org/dc/dcmitype/	http://dublincore.org/schemas/xmls/qdc/dcmitype.xsd

③ 将 DC 元素作为 XML 元素(建议全部用小写字母表示), DC 元素值作为对应的 XML 元素内容。如:

```
<dc:title>Dublin Core in XML</dc:title>
```

④ 对于重复的 DC 元素,在 XML 中也必须重复表示。例如,某篇文献有两个标题,名称分别是标题 A 和标题 B,则用 XML 表示为:

```
<dc:title>标题 A</dc:title>
```

```
<dc:title>标题 B</dc:title>
```

⑤ DC 的修饰词也应和其他 DC 元素同等对待。在此使用 DCMI 命名空间中与修饰词对应的名称。如:

```
<dcterms:available>2002-06</dcterms:available>
```

⑥ DC 的编码体系使用 XML 的“xsi:type”属性完成。例如:

```
<dc:identifier xsi:type="dcterms:URI">http://www.ukoln.ac.uk</dc:identifier>
```

⑦ 对于语种声明,使用“xml:lang”,例如:

```
<dc:subject xml:lang="en">seafood</dc:subject>
```

一个带修饰词的实例如下^①:

```
<?xml version="1.0"?>
```

```
<metadata
```

```
  xmlns="http://example.org/myapp/"
```

```
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
```

```
  xsi:schemaLocation=" http://example.org/myapp/ http://example.org/myapp/schema.xsd"
```

```
  xmlns:dc="http://purl.org/dc/elements/1.1/"
```

```
  xmlns:dcterms="http://purl.org/dc/terms/">
```

```
<dc:title>UKOLN</dc:title>
```

```
<dcterms:alternative> UK Office for Library and Information Networking
```

```
</dcterms:alternative>
```

```
<dc:subject>
```

national centre, network information support, library community, awareness, research, information services, public library networking, bibliographic management, distributed library systems, metadata, resource discovery, conferences, lectures, workshops

```
</dc:subject>
```

```
<dc:subject xsi:type="dcterms:DDC">062 </dc:subject>
```

```
<dc:subject xsi:type="dcterms:UDC"> 061(410) </dc:subject>
```

```

<dc:description>
    UKOLN is a national focus of expertise in digital information management. It provides policy,
    research and awareness services to the UK library, information and cultural heritage communities.
    UKOLN is based at the University of Bath.
</dc:description>
<dc:description xml:lang="fr">
    UKOLN est un centre national d'expertise dans la gestion de l'information digitale.
</dc:description>
<dc:publisher> UKOLN, University of Bath</dc:publisher>
<dc:isPartOf xsi:type="dcterms:URI"> http://www.bath.ac.uk/ </dc:isPartOf>
<dc:identifier xsi:type="dcterms:URI"> http://www.ukoln.ac.uk/ </dc:identifier>
<dc:modified xsi:type="dcterms:W3CDTF"> 2001-07-18</dc:modified>
<dc:format xsi:type="dcterms:IMT"> text/html </dc:format>
<dc:extent> 14 Kbytes </dc:extent>
</metadata>

```

3) 使用 RDF 描述

RDF 是用来描述资源及其关系的语言。例如，某文档资源的标题是 AAA，用 RDF 的图示法描述如图 4.6 所示。

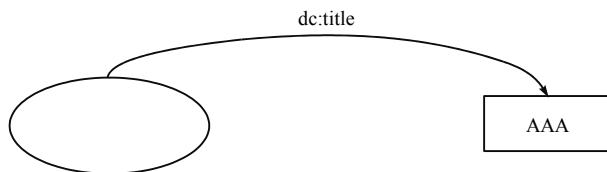


图 4.6 使用 RDF 图示的 DC 元素

其中，椭圆为该文档资源，矩形框中是表示标题的字符串值。

将图 4.7 用 XML 实现的代码如下：

```

<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description>
        <dc:title>AAA</dc:title>
    </rdf:Description>
</rdf:RDF>

```

其中，<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dc="http://purl.org/dc/elements/1.1/"> 分别用来说明 RDF 的命名空间和 DC 的命名空间（其中包含 DC 的元素定义）。

<dc:title>AAA</dc:title>描述资源的标题。

关于带修饰词的 DC 元素的 RDF 描述，在此不再讨论。

4.3.3 元数据描述框架

本书第 3 章第 3.4.4 节已讲述了 RDF（资源描述框架）的有关知识，RDF 作为一种元数

据描述框架, 本节主要从著录法角度举例说明 RDF 模型、语法和容器元素。

1. RDF 模型

RDF 数据模型为元数据的定义和使用提供了一个抽象的、概念化框架。基本的 RDF 数据模型由 3 类对象组成。

(1) 资源 (Resource)。由 RDF 描述的任何事物都可以称为资源, 如一本书、一个作者、一个网页、多个网页集合等。每个资源都有一个统一资源标识符 URI。

(2) 属性 (Properties)。也称性质, 用于描述某一资源的特定方面 (a Specific Aspect)、特征 (Characteristic)、属性 (Attribute) 和关系 (Relation)。每个属性都有其特定的含义, 定义其允许值、可描述的资源类型及与其他属性之间的关系。

(3) 陈述 (Statements)。也称声明, 由资源、资源上已定义的属性和该属性的值构成。陈述的这三个独立部分分别称为主体 (Subject)、谓词 (Predicate) 和客体 (Object)。其中客体 (即属性值) 可以是另一个资源或文字 (Literal)。

下面给出一个实例用于说明 RDF 模型。

例 4.1 Diane Hillmann 是资源 <http://dublincore.org/documents/2005/11/07/usageguide/> 的创建者。

该语句的 3 个部分如下。

(1) 主体 (Subject): 描述资源, 其 URI 是 <http://dublincore.org/documents/2005/11/07/usageguide/>。

(2) 谓词 (Predicate): 描述资源的属性, 在此是“创建者”。

(3) 客体 (Object): 表示属性值, 在此是指“创建者”的值, 是“Diane Hillmann”。

也可以使用以资源为节点的有向图方式显示。其中资源和属性值都是以节点表示的, 属性以有向弧表示, 如图 4.7 所示。

例 4.1 用图表示如图 4-10 所示。

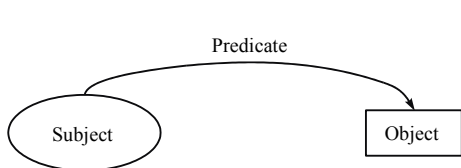


图 4.7 RDF 图示法



图 4.8 例 4.1 的简单节点和弧示意图

若客体本身也是一个资源 (或称结构化实体), 则客体也是以椭圆节点进行表示的。若该客体是匿名的, 则该节点为空节点, 若该客体有 URI, 则节点中给出 URI。

例 4.2 Diane Hillmann (E-mail 为 dih1@cornell.edu) 是资源 <http://dublincore.org/documents/2005/11/07/usageguide/> 的创建者。用图表示如图 4.9 所示。

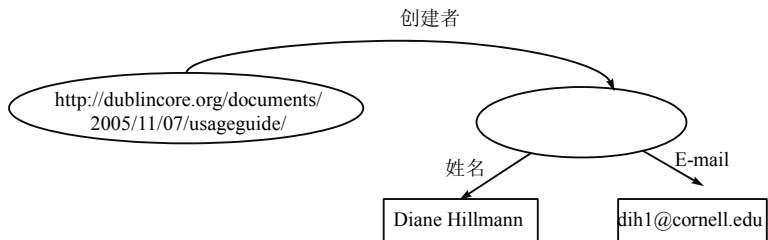


图 4.9 有结构化属性值

例 4.3 在例 4.2 中, 增加 Diane Hillmann 的信息所在

<http://www.cornell.edu/search/index.cfm?tab=people&netid=dih1&q=Diane%20Hillmann>, 则可以将图 4.9 改为如图 4.10 所示的形式。

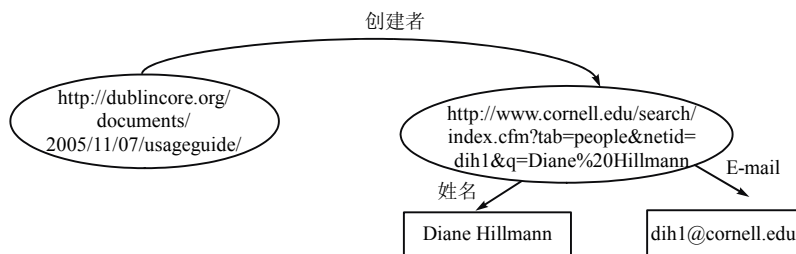


图 4.10 有标识符的结构化属性

2. 基于 XML 的 RDF 语法

RDF 语法主要用于元数据的创建和交换。RDF 数据模型有两类 XML 语法：序列化语法 (Serialization Syntax) 和简略语法 (Abbreviated Syntax)。序列化语法用非常规则的方式表达数据模型的全部功能；简略语法提供表达数据模型的子集的简洁形式^①。

1) 序列化语法

RDF 序列化语法形式如下^②：

[1] RDF	::= ['<rdf:RDF>'] description* ['</rdf:RDF>']
[2] description	::= '<rdf:Description' idAboutAttr? '>' propertyElt* '</rdf:Description>'
[3] idAboutAttr	::= idAttr aboutAttr
[4] aboutAttr	::= 'about=' URI-reference ''
[5] idAttr	::= 'ID=' IDsymbol ''
[6] propertyElt	::= '<' propName '>' value '</' propName '>' '<' propName resourceAttr '>'
[7] propName	::= QName
[8] value	::= description string
[9] resourceAttr	::= 'resource=' URI-reference ''
[10] QName	::= [NSprefix ':'] name
[11] URI-reference	::= string, interpreted per [URI]
[12] IDsymbol	::= (any legal XML name symbol)
[13] name	::= (any legal XML name symbol)
[14] NSprefix	::= (any legal XML namespace prefix)
[15] string	::= (any XML text, with "<", ">", and "&" escaped)

对语法的说明如下。

(1) RDF 元素作为文档的边界, 将一系列描述包裹起来。

(2) 一般而言, 一个资源同时具有若干个属性, 因此该语法将同一资源的陈述组合在 Description 元素中。Description 元素在 about 属性中为各个陈述应用的资源命名。如果没有 about 属性, 则表示该资源是一个新资源, 这样的资源可能是一个没有 URI 的实体资源的替代。

(3) Description 元素的 ID 属性表示行内 (In-line) 资源的标识。如果另一个 Description

① 张家耕, 谢晓竹. XML 网络编程技术[M]. 北京: 国防工业出版社, 2002.

② <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/> (2009-1-11).

元素或者属性值要引用这个行内资源,那么就在它自己的 `about` 属性中给出该 ID 的值。在同一个 `Description` 元素中不能同时出现 `about` 和 `ID` 属性。

(4) 一个 `Description` 元素可以有同一个属性的多个 `PropertyElt` 元素。每一个 `propertyElt` 元素在图中增加一条弧线。在 `propertyElt` 元素中, `resource` 属性指定了其他资源是该属性值,即客体是另一个 URI 表示的资源。

例 4.1 的语法如下:

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:s="http://description.org/schema/">
  <rdf:Description
    about="http://dublincore.org/documents/2005/11/07/usageguide/">
    <s:Creator>Diane Hillmann</s:Creator>
  </rdf:Description>
</rdf:RDF>
```

其中 `RDF` 是根,包裹了整个实例,其命名空间的声明是 `xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"`。有一个 `Description` 元素,其属性 `about` 对现有资源 `http://dublincore.org/documents/2005/11/07/usageguide/` 进行说明。前缀 `s` 表示一个特定的命名空间,在此其命名空间的声明是 `xmlns:s="http://description.org/schema/"`。元素 `Creator` 表示“创建者”,值为“Diane Hillmann”。

2) 简略语法

有 3 种形式的简略语法。第一种形式是针对没有重复属性的 `Description` 元素设计的,并且这些属性的值都是文字的。如上例就可以变为:

```
<rdf:RDF>
  <rdf:Description
    about="http://dublincore.org/documents/2005/11/07/usageguide/"
    s:Creator=Diane Hillmann/>
</rdf:RDF>
```

第二种形式是针对嵌套的 `Description` 元素。这种简写形式适用于这样的情况:客体是另一个资源,为第二个资源给的行内属性值都是字符串。此时,类似于将 XML 元素变为 XML 属性的转换是:将嵌套的 `Description` 元素的 `about` 属性作为 `propertyElt` 元素的 `Resource` 属性。

例 4.3 使用嵌套的 `Description` 元素的语法是:

```
<rdf:RDF>
  <rdf:Description about="http://dublincore.org/documents/2005/11/07/usageguide/">
    <s:Creator>
      <rdf:Description about="http://www.cornell.edu/search/index.cfm?tab=people&netid=dih1&q=Diane%20Hillmann">
        <v:Name>Diane Hillmann</v:Name>
        <v:Email>dih1@cornell.edu</v:Email>
      </rdf:Description>
    </s:Creator>
  </rdf:Description>
</rdf:RDF>
```

可以简化为：

```
<rdf:RDF>
  <rdf:Description about=" http://dublincore.org/documents/2005/11/07/usageguide/">
    <s:Creator rdf:resource=" http://www.cornell.edu/search/index.cfm?tab=people&netid=dih1
&q=Diane%20Hillmann "
      v:Name=" Diane Hillmann "
      v:Email=" dihl@cornell.edu " />
  </rdf:Description>
</rdf:RDF>
```

第三种简略方法适用于包含 Type 属性的 Description 元素。如例 4.3 中，Diane Hillmann 是一人名，再增加一个事实是：“http://www.cornell.edu/search/index.cfm?tab= people&netid= dihl&q=Diane%20Hillmann”表示 Person 元素的实例。用序列化语法则为：

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:s="http://description.org/schema/">
  <rdf:Description about=" http://dublincore.org/documents/2005/11/07/usageguide/">
    <s:Creator>
      <rdf:Description about=" http://www.cornell.edu/search/index.cfm?tab=people&netid=
dih1&q=Diane%20Hillmann ">
        <rdf:type resource="http://description.org/schema/Person"/>
        <v:Name> Diane Hillmann </v:Name>
        <v:Email> dihl@cornell.edu </v:Email>
      </rdf:Description>
    </s:Creator>
  </rdf:Description>
</rdf:RDF>
```

使用简略语法形式如下：

```
<rdf:RDF>
  <rdf:Description about=" http://dublincore.org/documents/2005/11/07/usageguide/">
    <s:Creator>
      <s:Person about=" http://www.cornell.edu/search/index.cfm?tab=people&netid=dih1
&q=Diane%20Hillmann ">
        <v:Name> Diane Hillmann </v:Name>
        <v:Email> dihl@cornell.edu </v:Email>
      </s:Person>
    </s:Creator>
  </rdf:Description>
</rdf:RDF>
```

3. 容器元素

当描述一组资源的集合（如作品的多个作者）时，需要使用 RDF 的容器。RDF 提供 3 种容器：Bag、Sequence 和 Alternative。在使用容器元素时，Type 属性用来指明容器对象的类型。容器中成员常用“rdf: _1”、“rdf: _2”等标签来说明。

1) 包 Bag

rdf:Bag 表示一个无序的资源或文字列表。常用来声明一个属性具有多个值, 值之间的顺序无关紧要, 允许重复值。例如, 选修某门课程的学生集合的描述就可以使用包。

2) 序列 Sequence

rdf:Seq 表示一个有序的资源或文字列表。常用来声明具有多值的属性, 值之间的顺序是非常重要的, 允许重复值。例如一篇文献的作者常区分为第一作者、第二作者等。

3) 替代 Alternative

rdf:Alt 表示一个属性的多个替代值表。例如文献题名的多语种翻译版本或一个资源的多个镜像站点。可以从替代值表中选取一个合适值。

RDF 容器语法如下^①:

[1] container	::= sequence bag alternative
[2] sequence	::= '<rdf:Seq' idAttr? '>' member* '</rdf:Seq>'
[3] bag	::= '<rdf:Bag' idAttr? '>' member* '</rdf:Bag>'
[4] alternative	::= '<rdf:Alt' idAttr? '>' member+ '</rdf:Alt>'
[5] member	::= referencedItem inlineItem
[6] referencedItem	::= '<rdf:li' resourceAttr '>'
[7] inlineItem	::= '<rdf:li>' value '</rdf:li>'

其中使用 **rdf:li** 元素替代 “**rdf:_1**”、“**rdf:_2**” 等标签, 避免显式使用数词。

例 4.4 华军软件园 (<http://www.newhua.com/>) 的镜像站点有: <http://86516.onlinedown.net/> 和 <http://sq.newhua.com/>。其 RDF 模型如图 4.11 所示。

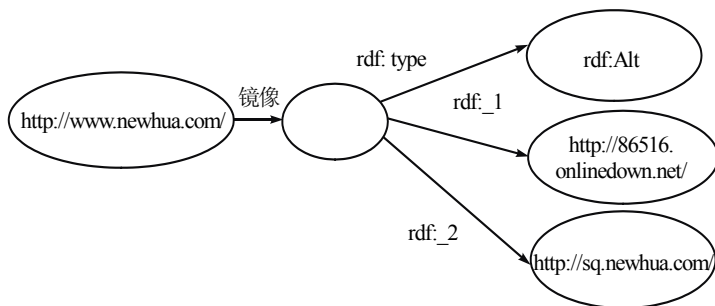


图 4.11 RDF 模型

其语法如下:

```

<rdf:RDF>
  <rdf:Description about=" http://www.newhua.com/">
    <s:DistributionSite>
      <rdf:Alt>
        <rdf:li resource=" http://86516.onlinedown.net/">
        <rdf:li resource=" http://sq.newhua.com/">
      </rdf:Alt>
    </s:DistributionSite>
  </rdf:Description>
</rdf:RDF>

```

① <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/> (2009-1-11)。

```
</rdf:Description>  
</rdf:RDF>
```

4.3.4 其他元数据

元数据标准有几十种，由于篇幅限制，在此仅介绍其中3种。

1. PICS

网络内容选择 PICS（Platform for Internet Content Selection）始于1995年8月，作为一种元数据标准，旨在描述 Internet 文档内容。特别是对敏感资料（如色情或暴力资料）的分级功能，用于帮助父母和老师控制儿童获取网络资源，过滤掉那些对自己或儿童不适合、不需要的内容。PICS 主要使用两种分级方式：一是使用标签实现自我分级，从而使内容供应商自愿对他们创建和发布的内容进行标签服务；二是使用第三方分级，制定自己的标签系统，从而使多个独立的标签服务利用附加标签与其他人创建和发布的内容联系起来。PICS 成员相信一个开放性的标签平台是一个保持 Internet 的活力和多样性的最好方式。两种分级方式可以使用户对接收到的内容在最大程度上进行控制，而不需要内容提供者的更多限制。PICS 技术规范最终在1996年年初完成，此后 PICS 被结合到一系列产品中，如基于 PICS 的分级服务和过滤工具被大量地开发出来^{①②}。

W3C 工作组就 PICS-1.1 已经完成如下4个技术方面的建议^{③④}。

（1）服务描述。定义描述分级服务的词汇和级别。

（2）标签格式和分布。定义了标签和标签传送方式的通用格式。其中标签传送主要有3种方式：在 HTML 文档中传送；使用 RFC-822 标头传送；与文档分开传送。

（3）PICS 规则。它是为过滤规则设置的相互交换格式。这些通过对使用 PICS 标签描述的 URLs 的允许或阻止规则可以很容易地安装或发送给搜索引擎。

（4）PICS 签名标签（Dsig）1.0 说明书。指定用 PICS 标签实现的数字签名的语法和语义。

此外，还有一个非官方但代表 PICS 工作组共识的关于默认和覆盖标签的推荐。

2. CDWA

艺术作品描述目录 CDWA（The Categories for the Description of Works of Art）是1996年由 AITF（the Art Information Task Force）开发的，主要为提供和使用艺术信息的团体或组织机构（如博物馆和档案馆）对艺术作品、建筑、其他物质文化作品等对象及其相关图像的描述及评价信息提供结构化概念框架。CDWA 共有 540 个大类及子类。在类目结构安排上，共有 31 个大类（具体见表 4.15），每一个大类结构由定义（DEFINITION）、子类（SUB CATEGORIES）和讨论（DISCUSSION）组成。每一个子类下有定义（DEFINITION）、实例（EXAMPLES）、讨论和指南（DISCUSSION and GUIDELINES）、术语/格式（TERMIN OLOGY/FORMAT）、相关类及评价（RELATED CATEGORIES and ACCESS）。

① 曹建. Internet 与 E-mail 安全防范实用技术[M]. 成都: 电子科技大学出版社, 1999.

② 各种元数据格式简介. <http://www.lib.hust.edu.cn/lib/dllib.nsf/1ce930115fcd2e7048256c9a0006f537/52809febff2fd74348256c9f002b29d5?OpenDocument> (2009-1-7).

③ Platform for Internet Content Selection (PICS). <http://www.w3.org/PICS/#Specs> (2009-01-07).

④ 李宏伟, 等. 网络地理信息系统与空间元数据[M]. 郑州: 黄河水利出版社, 2004.

表 4.15 CDWA 大类表^{①②③④}

类 名	中文译名	含 义	子 类
OBJECT/WORK	物件/作品	所描述作品的类型和数量	编目层次 (Catalog Level)、作品类型 (Object/Work Type)、该类型日期 (Object/Work Type Date)、组成部分 (Components/Parts)、备注 (Remarks)、引用 (Citations)
CLASSIFICATION	分类	在一个分类体系中, 依据相似特征将一件艺术或建筑作品归类	术语 (Classification Term)、备注 (Remarks)、引用 (Citations)
TITLES OR NAMES	题名	艺术、建筑等作品名称, 也包含作品题名的类型和时间	题名文本 (Title Text)、题名类型 (Title Type)、偏好 (Preference)、题名语种 (Title Language)、题名日期 (Title Date)、备注 (Remarks)、引用 (Citations)
CREATION	创作	艺术作品或建筑及其组件的创作、设计、执行或生产信息, 包括责任者、日期和地点	创作者 (Creator Description)、创作日期 (Creation Date)、创作地点 (Creation Place/Original Location)、作品文化 (Object/Work Culture)、委员 (Commissioner)、创作号 (Creation Numbers)、备注 (Remarks)、引用 (Citations)
STYLES/PERIODS/ GROUPS/MOVEMENTS	风格/时期/ 团体/运动	对于显现在作品中的特征描述, 包括作品的风格、历史时期、团体、画派或运动的描述	描述 (Styles/Periods Description)、标引词 (Styles/Periods Indexing Terms)、备注 (Remarks)、引用 (Citations)
MEASUREMENTS	测量	提供作品的规格、形状、比例和尺寸的信息	尺寸描述 (Dimensions Description)、类型 (Dimensions Type)、数值 (Dimensions Value)、单位 (Dimensions Unit)、程度 (Dimensions Extent)、比例类型 (Scale Type)、修饰 (Dimensions Qualifier)、日期 (Dimensions Date)、形状 (Shape)、格式/大小 (Format/Size)、备注 (Remarks)、引用 (Citations)
MATERIALS /TECHNIQUES	材料/技术	作品创作中使用的材料物质及生产制造技术、过程或方法	描述 (Materials/Techniques Description)、标识 (Materials/Techniques Flag)、程度 (Materials/Techniques Extent)、作用 (Material/Techniques Role)、名称 (Materials/Technique Name)、材料颜色 (Material Color)、材料原产地 (Material Source Place)、水印 (Watermarks)、活动 (Performance Actions)、备注 (Remarks)、引用 (Citations)
INSCRIPTIONS/ MARKS	题刻/标记	对镶嵌、贴、写、刻或附著于作品之上的记号、字符、注释、文本或标签部分的描述	题词或描述 (Inscription Transcription or Description)、类型 (Inscription Type)、作者 (Inscription Author)、位置 (Inscription Location)、语种 (Inscription Language)、字体/字形 (Typeface/Letterform)、标记识别 (Mark Identification)、日期 (Inscription Date)、备注 (Remarks)、引用 (Citations)
STATE	阶段	作品创作阶段描述	阶段描述 (State Description)、识别 (State Identification)、已知阶段 (Known States)、备注 (Remarks)、引用 (Citations)

① http://metadata.teldap.tw/standard/standard-big5/cdwa_2-0_draft.pdf (2009-1-6)。② http://www.getty.edu/research/conducting_research/standards/cdwa/ (2009-1-8)。③ http://www.getty.edu/research/publications/electronic_publications/cdwa/definitions.html (2013-12-23)

④ 肖婷.应用 CDWA 标准描述数字宋画作品的探索.图书情报工作, 2011(9):101-146

续表

类 名	中 文 译 名	含 义	子 类
EDITION	版本	用来确定一组同时发行的作品中的某一作品, 或者定义一件作品的发行的先后版次	版本描述 (Edition Description)、版次或名称 (Edition Number or Name)、版印序号 (Impression Number)、版本印数 (Edition Size)、备注 (Remarks)、引用 (Citations)
FACTURE	制作手法	作品制作方法的细节描述	描述 (Facture Description)、备注 (Remarks)、引用 (Citations)
ORIENTATION/ ARRANGEMENT	方位/布置	作品展示的方式	描述 (Orientation/Arrangement Description)、标引词 (Orientation Indexing Terms)、备注 (Remarks)、引用 (Citations)
PHYSICAL DESCRIPTION	形式描述	以一般性术语描述作品的外观, 而不涉及主题。包括作品装饰中的图案、纹饰等	形式外观 (Physical Appearance)、描述标引词 (Physical Description Indexing Terms)、备注 (Remarks)、引用 (Citations)
CONDITION/ EXAMINATION HISTORY	现状/鉴定 历史	在特定时期对作品的物理状态、特征和完整性的评估	描述 (Condition/Examination Description)、类型 (Examination Type)、鉴定人 (Examination Agent)、日期 (Examination Date)、地点 (Examination Place)、备注 (Remarks)、引用 (Citations)
CONSERVATION /TREATMENT HISTORY	保存/处理 历史	作品曾经历过的修复、保存的程序或行为	描述 (Conservation/Treatment Description)、处理类型 (Treatment Type)、处理人 (Treatment Agent)、处理时间 (Treatment Date)、地点 (Treatment Place)、备注 (Remarks)、引用 (Citations)
SUBJECT MATTER	主题	作品所描述的主题。在物件或建筑中若无描述性内容, 则可涵盖其功能	描述 (Subject Display)、概要主题词 (General Subject Terms)、详细主题词 (Specific Subject Terms)、其他术语 (Outside Iconography Terms)、诠释史 (Subject Interpretive History)、备注 (Remarks)、引用 (Citations)
CONTEXT	背景 (脉络)	作品创作或存在过程中, 与其相关的政治、社会、经济或宗教事件或运动	历史/文化事件 (Historical/Cultural Events)、建筑背景 (Architectural Context)、考古背景 (Archaeological Context)、历史背景场所 (Historical Location Context)、备注 (Remarks)、引用 (Citations)
DESCRIPTIVE NOTE	描述性注释	对作品的文字描述	描述性注释文本 (Descriptive Note Text)、备注 (Remarks)、引用 (Citations)
CRITICAL RESPONSES	评论	艺术家、建筑家、艺术史家、艺术评论家、艺术从业者等人对作品的评论意见	评述 (Critical Comment)、评述文档类型 (Comment Document Type)、评论人 (Comment Author)、评论日期 (Comment Date)、评论环境 (Comment Circumstances)、备注 (Remarks)、引用 (Citations)
RELATED WORKS	相关作品	描述与该作品相关的作品及它们之间的关系	标签/标识 (Related Work Label/Identification)、更宽泛的背景 (Work Broader Context)、关联号 (Relationship Number)、备注 (Remarks)、引用 (Citations)
CURRENT LOCATION	现藏地点	现在收藏该作品的地点和地理位置	描述 (Current Location Description)、典藏位置 (Current Repository/Geographic Location)、标签/标识 (Object/Work Label/Identification)、备注 (Remarks)、引用 (Citations)

续表

类 名	中 文 译 名	含 义	子 类
COPYRIGHT/ RESTRICTIONS	版权/限制	拥有作品的使用、展览或复制作品权限的个人或团体的识别及对这些权限的限制	版权声明 (Copyright Statement)、版权所有者的姓名 (Copyright Holder Name)、地点 (Copyright Place)、时间 (Copyright Date)、备注 (Remarks)、引用 (Citations)
OWNERSHIP/ COLLECTING HISTORY	所有权/收藏历史	作品从创作至今的拥有者的起源或历史	起源描述 (Provenance Description)、转让模式 (Transfer Mode)、成本或价值 (Cost or Value)、法定地位 (Legal Status)、所有者 (Owner/Agent)、所有权地点 (Ownership Place)、时间 (Ownership Date)、所有者号码 (Owner's Numbers)、来源附注 (Owner's Credit Line)、备注 (Remarks)、引用 (Citations)
EXHIBITION/ LOAN HISTORY	展览/借出历史	作品公开展示的历史记录	描述 (Exhibition/Loan Description)、展览名 (Exhibition Title or Name)、展览类型 (Exhibition Type)、展览管理者 (Exhibition Curator)、组织者 (Exhibition Organizer)、赞助人 (Exhibition Sponsor)、展出场所 (Exhibition Venue)、展出对象号码 (Exhibition Object Number)、展出作品标签/标识 (Exhibition Object/Work Label/Identification)、备注 (Remarks)、引用 (Citations)
CATALOGING HISTORY	编目历史	对作品的描述的产生和修改文档, 包括描述者、描述时间及相关注释	编目机构 (Cataloging Institution)、编目者姓名 (Cataloger Name)、编目者操作行为 (Cataloger Action)、编目者操作影响的记录区域 (Area of Record Affected)、编目日期 (Cataloging Date)、备注 (Remarks)、作品记录 ID (Object/Work Record ID)
RELATED VISUAL DOCUM ENTATION	相关的视觉文档	描述作品的影像信息	图像参考 (Image References)、图像标签/标识 (Image Label/Identification)
RELATED TEXTUAL REFERENCES	相关的参考文献	指对作品被描述的文本信息源的引用文献, 包括出版的书目资料、网站、档案文献、未出版的手稿及学者或主题专家的口头意见等	引文 (Citations for Sources)、简略引文 (Source Brief Citation)、备注 (Remarks)、引文规范记录 ID (Citations Authority Record ID)
PERSON/CORP ORATE BODY AUTHORITY	个人/团体规范	作品设计和制作的责任者, 包括艺术家、建筑家和其他个人和团体信息	个人规范记录类型 (Person Authority Record Type)、名称 (Person Name)、自传 (Display Biography)、出生时间 (Birth Date)、死亡时间 (Death Date)、出生地 (Birth Place)、死亡地点 (Death Place)、个人国籍 (Person Nationality/Culture/Race)、性别 (Gender)、职业 (Life Roles)、个人/团体事件 (Person/Corporate Body Event)、相关个人/团体 (Related Person/Corporate Body)、隶属于个人/团体 (Person/

续表

类 名	中 文 译 名	含 义	子 类
PERSON/CORPORATE BODY AUTHORITY	个人/团体规范	作品设计和制作的责任者，包括艺术家、建筑家和其他个人和团体信息	Corporate Body Broader Context）、个人/团体标签/标识（Person/Corporate Body Label/Identification）、个人/团体描述性注释（Person/Corporate Body Descriptive Note）、备注（Remarks）、引用（Citations）、个人规范记录 ID（Person Authority Record ID）
PLACE/LOCATION AUTHORITY	地区/位置规范	作品或创作者相关的地理位置信息规范	位置规范记录类型（Place Authority Record Type）、位置名（Place Name）、地理坐标（Geographic Coordinates）、位置类型（Place Types）、相关位置（Related Places）、更宽泛的位置（Place Broader Context）、位置标签/标识（Place/Location Label/Identification）、位置描述性注释（Place/Location Descriptive Note）、备注（Remarks）、引用（Citations）、位置规范记录 ID（Place Authority Record ID）
GENERIC CONCEPT AUTHORITY	一般概念规范	在编目或描述作品中需要的一般概念的信息规范	概念规范记录类型（Concept Authority Record Type）、术语（Generic Concept Term）、相关一般概念（Related Generic Concepts）、上位概念（Concept Broader Context）、一般概念标签/标识（Generic Concept Label/Identification）、概念范围注释（Concept Scope Note）、备注（Remarks）、引用（Citations）、概念规范记录 ID（Concept Authority Record ID）
SUBJECT AUTHORITY	主题规范	有名称的图像、文字、动物、故事或事件等信息规范	主题规范记录类型（Subject Authority Record Type）、主题名称（Subject Name）、主题日期（Subject Date）、主题作用/属性（Subject Roles/Attributes）、相关主题（Related Subject）、上位类主题（Subject Broader Context）、相关位置（Related Place/Location）、相关人/法人团体（Related Person/Corporate Body）、相关通用概念（Related Generic Concept）、主题标签/标识（Subject Label/Identification）、主题描述性注释（Subject Descriptive Note）、备注（Remarks）、引用（Citations）、主题规范记录 ID（Subject Authority Record ID）

3. FGDC（联邦地理数据委员会）

联邦地理数据委员会 FGDC（the Federal Geographic Data Committee）于 1992 年 6 月赞助召开空间数据的信息交换论坛（Information Exchange Forum on Spatial Data）。在该次论坛上讨论了元数据的标准化、使用和元数据系统工具。并开始起草元数据标准方案，从 1992 年 10 月到 1993 年 4 月就该草案进行公开审查与修订。并在 1994 年 6 月 8 日通过数字化地理元数据的内容标准（CSDGM, Content Standards for Digital Geospatial Metadata,），但通常仍叫做 FGDC。1994 年 4 月 11 日，美国总统克林顿签发的 12906 号总统令“地理数据的获取与存储：国家地理空间数据基础设施”（Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure）中第三部分（b）段中要求联邦代理机构从 1995 年开始使用这个标准来为新的地理空间数据编写文档。1997 年完成第 2 版并开始审查，1998 年认可第 2 版（FGDC-STD-001-1998）。

FGDC 标准有 11 部分（从 Section 0 到 Section 10），每一部分提供了复合元素及其定义、元素值类型、元素是否是必备性或可重复性说明。其中，Section 0 “元数据”是起始节点，它由该标准的主要部分组成；Section 1 到 Section 7 是该标准的主要部分；Section 8 到 Section

10 是附加部分, 这 3 部分不能单独使用。其结构如图 4.12 所示^①。

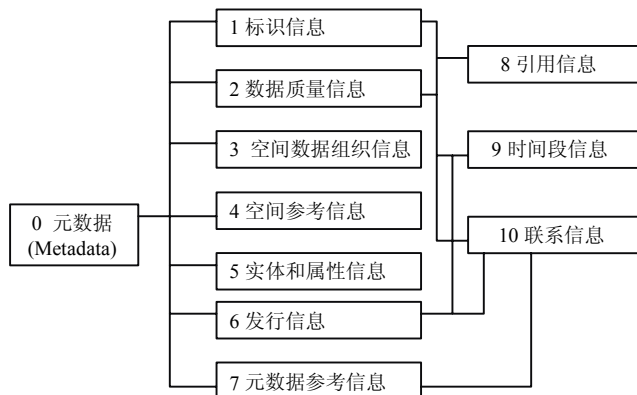


图 4.12 FGDC 标准结构图

这七个主要部分集中描述了地理空间数据的不同方面的特征^{②③④⑤}。

(1) 标识信息 (Identification Information)。关于数据集的基本信息, 其类型是复合型 (Type: compound), 简称 (Short Name) 是 Idinfo。包含元素: 引用 (Citation)、描述 (Description)、内容时间范围 (Time_Period_of_Content)、状态 (Status)、空间域 (Spatial_Domain)、关键词 (Keywords)、存取限制 (Access_Constraints)、使用限制 (Use_Constraints)、联系点 (Point_of_Contact)、浏览图像 (Browse_Graphic)、数据集贡献度认可 (Data_Set_Credit)、安全信息 (Security_Information)、原始数据集环境 (Native_Data_Set_Environment)、交叉参照 (Cross_Reference)。

(2) 数据质量信息 (Data Quality Information)。关于数据集质量的综合评价的信息, 其类型是复合型 (Type: Compound), 简称 (Short Name) 是 Dataqual。包含元素: 属性准确性 (Attribute_Accuracy)、逻辑一致性报告 (Logical_Consistency_Report)、完整性报告 (Completeness_Report)、位置准确性 (Positional_Accuracy)、谱系 (Lineage)、云层覆盖 (Cloud_Cover)。

(3) 空间数据组织信息 (Spatial Data Organization Information)。说明用于在数据集中表示空间信息的机理。其类型是复合型 (Type: Compound), 简称 (Short Name) 是 Spdoinfo。包含元素: 间接空间参照 (Indirect_Spatial_Reference)、直接空间参照 (Direct_Spatial_Reference_Method)、点对象和矢量对象信息 (Point_and_Vector_Object_Information)、光栅对象信息 (Raster_Object_Information)。

(4) 空间参考信息 (Spatial Reference Information)。数据集中坐标参考框架的描述, 其类型是复合型 (Type: compound), 简称 (Short Name) 是 Spref。包含元素: 水平坐标体系定义 (Horizontal_Coordinate_System_Definition)、垂直坐标体系定义 (Vertical_Coordinate_System_Definition)。

① http://www.fgdc.gov/metadata/documents/workbook_0501_bmk.pdf (2009-1-8)。

② 总装备部. 卫星应用现状与发展 (上册) [M]. 北京: 中国科学技术出版社, 2001.

③ 史文中. 空间数据与空间分析不确定性原理 [M]. 北京: 科学出版社, 2005.

④ http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf (2009-1-8)。

⑤ 张晓林. 元数据研究与应用 [M]. 北京: 北京图书馆出版社, 2002.

(5) 实体和属性信息 (Entity and Attribute Information)。关于数据集内容的信息, 包括实体类型及其属性及属性的域值; 其类型是复合型 (Type: Compound), 简称 (Short Name) 是 Eainfo。包含元素: 详细描述 (Detailed_Description)、总体描述 (Overview_Description)。

(6) 发行信息 (Distribution Information)。关于数据获取方式选择、发行者的信息, 其类型是复合型 (Type: Compound), 简称 (Short Name) 是 Distinfo。包含元素: 发行者 (Distributor)、资源描述 (Resource_Description)、发行责任 (Distribution_Liability)、标准订购程序 (Standard_Order_Process)、定制订购程序 (Custom_Order_Process)、技术前提 (Technical_Prerequisites)、可用时限 (Available_Time_Period)。

(7) 元数据参考信息 (Metadata Reference Information)。关于元数据信息的新颖性 (Currentness) 及责任者信息, 其类型是复合型 (Type: Compound), 简称 (Short Name) 是 Metainfo。包含元素: 元数据日期 (Metadata_Date)、元数据审查日期 (Metadata_Review_Date)、元数据未来审查日期 (Metadata_Future_Review_Date)、元数据联系 (Metadata_Contact)、元数据标准名称 (Metadata_Standard_Name)、元数据标准版本 (Metadata_Standard_Version)、元数据时间转换 (Metadata_Time_Convention)、元数据获取限制 (Metadata_Access_Constraints)、元数据使用限制 (Metadata_Use_Constraints)、元数据安全信息 (Metadata_Security_Information)、元数据扩展 (Metadata_Extensions)。

其中, 标识信息和元数据参考信息是必备的, 其他 5 个则是若有则必备。

FGDC 的 3 个附加部分分别简述如下。

(1) 引用信息 (Citation Information)。指被用于数据集的推荐参考。其类型是复合型 (Type: Compound), 简称 (Short Name) 是 Citeinfo。包含元素: 创建者 (Originator)、发表日期 (Publication_Date)、发表时间 (Publication_Time)、标题 (Title)、版本 (Edition)、地理空间数据表现方式 (Geospatial_Data_Presentation_Form)、序列信息 (Series_Information)、发表信息 (Publication_Information)、其他引用细节 (Other_Citation_Details)、在线链接 (Online_Linkage)、上层对象引用信息 (Larger_Work_Citation)。

(2) 时间段信息 (Time Period Information)。关于时间的日期时间信息。其类型是复合型 (Type: Compound), 简称 (Short Name) 是 Timeinfo。包含元素: 单一日期/时间 (Single_Date/Time)、多日期/时间 (Multiple_Dates/Times)、日期/时间范围 (Range_of_Dates/Times)。

(3) 联系信息 (Contact Information)。与数据集有关的个人、组织联系方式和身份。其类型是复合型 (Type: Compound), 简称 (Short Name) 是 Cntinfo。包含元素: 主要联系人 (Contact_Person_Primary)、主要联系组织 (Contact_Organization_Primary)、联系人地位 (Contact_Position)、联系地址 (Contact_Address)、联系电话 (Contact_Voice_Telephone)、听力受损者联系电话 (Contact_TDD/TTY_Telephone)、传真号 (Contact_Facsimile_Telephone)、E-mail 地址 (Contact_Electronic_Mail_Address)、服务时间段 (Hours_of_Service)、联系指导 (Contact_Instructions)。



本章小结

信息著录是指为揭示、报道与检索的目的, 依据特定的规则与方法, 对信息资源的形式特征特征与内容特征进行描述、标引并使其有序化的方法和过程。在这一过程中, 所依据的规则和方法是信息著录质量的重要保证之一。本章选取国内外应用比较广泛的具有代表性的

文献著录规则,进行了较为全面的介绍,同时介绍了 MARC 21 记录格式,以及元数据的定义、功能,并以 DC、CDWA 等为例介绍其元素组成和具体应用。



问题讨论

1. 简述《国际标准书目著录》(ISBD)和《英美编目条例》(第二版)(AACR 2)的编制体例和主要特点。
2. 试比较信息著录中“款目”与“记录”两个概念的异同。
3. 《文献著录总则》包括哪些著录项目?
4. 简述 MARC 记录的基本格式。
5. 简述 MARC 的优缺点。
6. 简述元数据的功能。
7. 简述 DC 的元素及其含义。



第 5 章

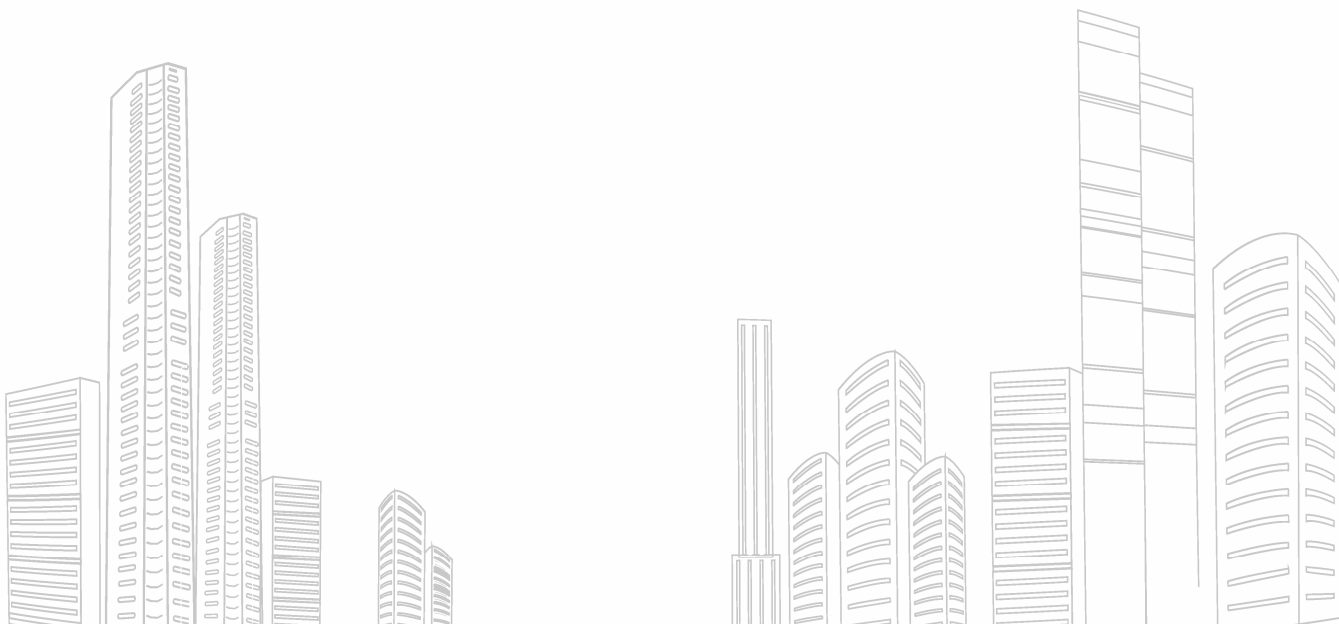
信息标引法

本章引言

标引是通过对文献或信息资源的分析,选用确切的检索标识,如类号、主题词、关键词、人名、地名等,用以反映该文献或资源内容的过程。利用何种检索标识进行标引,就形成了各种标引法。按使用检索标识或语言的类型划分,可有分类标引法、主题标引法、关键词标引法等,由于计算机信息检索系统和信息数据库的应用,又出现了自动标引。本章对分类标引、主题标引、自动分类与自动标引等进行了详细论述。

本章重点

- 分类标引方法;
- 分类标引规则;
- 主题标引的方法;
- 关键词语言的原理与类型;
- 自动标引技术。



5.1 分类标引

5.1.1 分类标引要求

标引是通过对文献或信息资源的分析,选用确切的检索标识,如类号、主题词、关键词、人名、地名等,用以反映该文献或资源内容的过程。通常指选用检索语言词或自然语言词反映文献内容,标引是内容的主题分析和用词表达两个步骤的结合。英文 Indexing 一词,意为标引(另一指索引法)。标引是文献加工中的重要环节,是款目或记录编排的基础和根据,对信息检索效果有直接的、决定性的影响。通过标引,各种目录、索引等检索工具才能编成。各种标引词存储于计算机内,才能实现文献或信息的检索。利用何种检索标识进行标引,就形成了各种标引法。按使用检索标识或语言的类型划分,可有分类标引法、主题标引法、关键词标引法等。由于计算机信息检索系统和信息数据库的应用,又出现了自动标引。分类标引又称为归类,是指依据一定的分类语言,对信息资源的内容特征进行分析、判断、选择,赋予分类标识的过程。

分类标引工作是对信息资源进行分类组织的基础和前提,对信息资源的开发利用具有重要的意义。通过对信息资源赋予分类标识,信息机构就可以将各种资源纳入相应的知识门类,建立起相应的分类检索系统。这样,用户可以根据一定的资源特征,按照系统提供的途径进行查找,进而从资源集合中检索出所需要的文献。同时,尽管有各种组织文献的方法,但分类标引依然是多数文献单位用来组织文献的依据,与文献单位各项工作的开展密切相关。

为了保证信息资源分类的质量,分类标引工作必须遵循以下要求。

1. 准确

准确,主要具有两层含义:一是归类正确,即将信息资源归入与其内容相对应的学科和专业;二是归类确切,即将信息资源归入分类体系中最专指、最切合其内容的类目。这就需要标引者不仅对信息资源的内容有一定的分析和判断能力,而且还要了解类目体系。

2. 充分

充分,指标引者能根据用户使用的需要,充分揭示有检索价值的信息资源的主题。一般来说,只讨论一个主题对象的信息资源,只归入一个对应的类目;同时讨论两个或三个主题的信息资源,则同时对这两个或三个主题对象进行分类标引,以便用户可以从不同论题出发检索出该资源,提高查全率。当然,对信息资源内容的揭示也应有所控制,过深的标引会降低查准率。

3. 一致

一致,指对同一主题内容资源的标引结果应保证一致,提高系统的查全率和查准率,这是检索系统提高检索效率的基本要求。为此,不仅应提高标引人员的素养,而且还要规范标引程序、建立明确的标引规则。

4. 适用

分类标引时应兼顾系统的特点和用户的检索需要,使标引结果适合使用。

5.1.2 分类标引方法

1. 类目辨析

只有掌握类目辨析的方法,准确了解类目的含义和范围,才能按照信息资源的内容和特

征将其归入分类体系中的相应类目。在分类法中,类目体系既要按照类目之间的等级和并列等形式进行排列,又要使用多种手段揭示类目之间的各种横向联系。因此,要弄清类目包含的范围,除了通过类目名称和有关的说明性文字进行了解外,还应善于根据类表的特点,从多个角度进行辨析。

具体来说,类目辨析包括以下几个方面的内容。

(1) 根据上位类与下位类的关系了解类目的含义。

上位类与下位类的关系是分类体系中基本的关系类型,反映了类目之间的纵向关系,彼此互为作用,联系紧密。通过对类目进行层层划分而建立起类目体系,而且,在类目展开的过程中,上位类对下位类的含义起着限定的作用。一般来说,在分类法中,为了使类目名称尽可能简短,下位类的类名往往省略其上位类的含义,只列出子概念的名称。因此,在使用下位类分类标引时应注意结合其上位类的含义加以理解。下面是《中图法》中部分类目的实例。

.....

J72 中国舞蹈、舞剧

J721 舞蹈图谱

J721.1 古代

J721.7 现代

.....

J722 各种舞蹈

.....

J723 各种舞剧

.....

其中,“舞蹈图谱”是“中国舞蹈、舞剧”的下位类,它的实际含义是“中国舞蹈图谱”。同样,“古代”和“现代”均为“舞蹈图谱”的下位类,它们的实际含义分别是“中国古代舞蹈图谱”和“中国现代舞蹈图谱”。

同样,网络分类体系中,某一具体类目的含义也必须结合其上位类的含义加以理解。例如,在搜狐分类目录中,“工商经济”下的“政策、法规”和“组织机构”类的实际含义分别是“工商经济政策、法规”和“工商经济组织机构”。

反过来,下位类对上位类的范围也具有揭示作用。也就是说,要了解一个类目所包括的范围,如果该类具有下位类,就可以由它的下位类来间接地确定它大致的外延。例如,“J72 中国舞蹈、舞剧”的下位类“舞蹈图谱”、“各种舞蹈”、“各种舞剧”,也可以反过来说明它们的上位类所包括的具体的对象和范围。

(2) 根据同位类之间的关系了解类目的含义。

在分类法中,同位类的设置通常是遵循逻辑划分规则进行的,一般情况下,采用某一个分类标准对某一个上位类进行划分而得到一组同位类,同位类之间的关系是相互排斥的,彼此界限比较明确。但在实际应用中,根据需要,有时也会同时采用两个或多个不同的划分标准对同一个类目进行一次性划分,这样,就会使部分同位类之间出现概念相互交叉的情况。对此,一般应根据同位类的特点,对相关类目的含义和范围加以限定,明确区分的界限。例如:

U448. 1/. 5 各种桥梁均可仿 U441/ U447 分。例如,《铁路桥梁的水文勘测》
入 U448. 132. 3。如遇多主题因素的桥梁文献,入编列在前的
类。例如,《铁路钢筋混凝土斜拉桥》入 U448. 13。

U448. 1 桥梁:按用途分

U448. 11 人行桥

U448. 12 两用桥

.....	
U448. 2	桥梁：按结构分
U448. 21	梁式桥
U448. 22	拱式桥
.....	
U448. 3	桥梁：按材料分
U448. 31	木桥、竹桥
U448. 32	石桥、砖桥
.....	
U448. 4	桥梁：按形式分
U448. 41	斜桥
U448. 42	曲弧桥、弯道桥
.....	
U448. 5	桥梁：按桥面系位置分
U448. 51	上承式桥
U448. 52	中承式桥
U448. 53	下承式桥
.....	

上例中，按不同标准区分出来的各个类目之间在包含的范围上存在着不同程度的交叉。所以，在确定特定信息资源的归属时，必须明确相关类目之间的关系，以便正确归类。但在网络分类体系中，从不同角度设置的同位类有时是以多维方式揭示的。

(3) 根据类目注释了解相关类目的含义和范围。

各个学科知识之间具有错综复杂的关系，反映到类目体系中，类目之间也具有纵横交错的关系。很多情况下，一个主题内容同时与多个知识门类具有联系，给明确判断其归属带来了困难。为了解决这类问题，分类表通常以注释的方式规定相关类目之间的区分界限，这些注释对于保证分类标引的准确性至关重要。分类标引时应注意通过这些注释了解类目的含义和范围，确切归类。例如，《中图法》中：

O655.1 重量分析

电重量分析入 O657.12；有机重量分析入 O656.33；

电解析法入 O657.13。

R249.1 医案、医话汇编

诸家医案分类汇编和合刻入此。

除此以外，对一些特殊的设类方法，类表一般通过注释的方式对类目的使用方法加以规定，借以明确类目的范围。例如，前面所列举的《中图法》“U448 各种桥梁”类在采用多个标准平行设类的同时，在类下注明：“如遇多主题因素的桥梁文献，入前面编列的类。”这实质上也是对有关类目包含范围的一种限定。

(4) 应按照类目体系展开的规律了解类目的含义和范围。

了解类目的含义和范围是分类标引的基础和保证，而分类法中类目的含义和范围与类目体系展开的方式密切联系。除了上述几个方面外，掌握类目展开的规律，也是正确了解类目的含义和范围的必要途径。掌握类目展开的规律，需要对分类体系进行全面的了解。

目前使用最为广泛的传统等级列举式分类法的类目体系，是一个按照从总到分、从一般到专门的方式，层层展开的线性系统。采用这种分类法标引时，就必须在掌握类目体系展开

的这些规律的基础上,了解类目的确切含义和范围,准确把握不同层次相关类目之间的关系,才能实现确切归类的目标。

以《中图法》为例,其类目体系是按学科为中心展开的,不同学科在类目体系展开时往往具有自身的特点。例如,社会科学范畴的学科,在类目体系展开时,往往十分重视地区的划分,按照“世界→中国→各国”的顺序依次排列类目。在地区概念下,基本上依然按照从总到分、从一般到专门的方式展开类目体系。因此,需要正确辨析属于不同地区范围的特定类目的含义。例如,“英美文学史”应标引为“I109”;“英法文学史”应标引为“I560.09”。因为,“英美”分处两个洲,属于“世界”范围,而“英法”同处一洲,属于“各国”范围。

网络分类法的类目体系总体上也是采用层层划分的方式展开的,但与传统文献分类法不同的是,网络分类法往往采用按照主题集中信息资源的方式来设置类目,通过超文本链接的形式对分属于不同门类的相同类目加以重复反映。所以,对相关内容的归属及类目的范围的了解,必须结合该分类体系的展开特点进行。

了解类目含义和范围还涉及许多因素,一般情况下,只要了解类目之间的各种相关联系,掌握类表展开的主要规律,就基本可以准确地辨析类目了。

2. 号码配置

号码配置是指在对信息资源进行分类标引的过程中,在按照信息资源主题内容归类的同时,根据分类法的要求获取相应分类号码的操作。分类号码即分类标记,是类目的代号,是信息资源进行分类组织的依据。正确进行号码配置是分类标引者必须掌握的基本技能。

一般情况下,号码配置的方法与所采用的分类工具及其标记特点密切联系。下面以《中图法》为例,简单介绍传统文献分类法的号码配置规律和方法。

《中图法》的号码配置主要有以下两种情况。

第一种情况是,单独且一次性使用主表就可以获得表达特定信息资源的完整的分类号码。

这种分类标引的方法比较简单,通常是根据特定信息资源的内容和特点,在层层展开的类目体系中,采用层层区分的方法,正确判断该信息资源在分类体系中的确切位置,直接获得完整的分类号码。例如,对“中国教育改革评价”这一主题进行分类标引时,可根据其所属的学科,在基本大类“G 文化、科学、教育、体育”下的“教育”中进行查找,直接得到该主题的分类号码是:“G521”。

第二种情况是,必须将不同成分的号码进行组配标引才能获得完整的分类号码。

结合组配手段进行号码配置,通常需要根据类表的规定,将表中代表不同成分的号码组合在一起。具体来说,《中图法》的组配标引包括以下三种类型。

(1) 使用复分表。在使用主表类目的同时,按照类表的要求结合使用复分表的类目进一步区分,即将主表号码与复分表号码进行组配,获得完整的分类号码。例如,对主题“陕西出土文物图录”标引时,可在主表类目“K873 出土文物图录”后加上中国地区表中陕西的号码“41”,得到该主题的号码为:K873.41。

(2) 仿分。利用仿分类目下同类性质的子目进一步细分。仿分时,一般在需要仿分的类号之后加上仿分号码即可。例如,对主题“我国城市公交经营管理问题研究”标引时,可在表示“中国城市交通运输经济”的号码“F572”后,加上从“城市交通运输经济理论”的号码“F570”后子目中表示“公共汽车”的号码“71”,得到该主题的号码为:F572.71。

(3) 类间组配。按照分类法的要求,结合使用特定的辅助符号,将一个主类号与其他与文献内容相关的主类号加以组合,用以表达信息资源的主题内容。例如,对主题“医学书目”标引时,可通过组配符号“:”将表示专科目录的类号“Z88”与表示“医学”的类号“R”组配联结,得到该主题的号码为:Z88:R。

以《中图法》作为分类标引工具进行号码组配时,需要注意以下问题。

(1) 删除含义相重的号码问题。

号码组配时,如果复分或仿分的号码与主表号码的成分含义相重,应删除重号。

例如,对主题“日本文物考古”标引时,按照类表的要求,应于“K883/887 各国文物考古”类下按世界地区表分。世界地区表中日本的号码为“313”,其中,“3”表示亚洲,其含义与“K883/887”中的“3/7”中的“3”相重,删去重号后,得到该主题的号码为:K883.13。

(2) 组配次序问题。

号码的组配次序,通常应按照类表注释指定的顺序进行。凡于某类下注明使用复分表配号,该复分号即应配于该主类号之后;当复分超过一次时,各次复分号码均应按注释规定的次序配号。

如“美国诗歌研究”的配号次序依次为:各国文学“I3/7”、美国“712”、“072 诗歌(的评论和研究)”。其中,“I3/7”中的“3/7”表示各国的号码范围,与“712”中的“7”南北美洲的含义相重,删去重号后,得到该主题的号码为:I712.072。

(3) 关于复分、仿分中号码加“0”问题。

《中图法》的标记符号是一个按照类目的次序和等级展开的符号系统。为了避免在复分、仿分配号时出现与类目体系中已有号码相重或与原有排列原则不一致的现象,类表规定通过在复分、仿分时加“0”来表示新类目的插入。《中图法》规定的复分、仿分加“0”的要点如下。

① 社会科学各类中,凡具有概括性地区属性的类目,如再依据其他标准复分或仿分时,组号时均须在主类号后先加“0”再复分或仿分。例如,“西欧财政史”应标引为:F815.609。

② 社会科学各类中的各级上位类,如果与其下位类之间采用层累标记制编号,当该上位类再依其他标准复分或仿分时,均需要在主类号后先加“0”再复分或仿分。例如,“奥运会滑雪成绩”应标引为:G863.108.112。

③ 在社会科学各类,凡属于越级复分的,均应在最后复分依据的复分子目号前加“0”。例如,“中国骑兵的教育与训练”应标引为:E271.203。

④ 历史大类下中国各代史仿通史分时,须加“0”。例如,“盛唐史料”应标引为:K242.206。

⑤ 自然科学各类仿“一般性问题”分时,须加“0”。例如,“电子数字计算机的检修与维护”应标引为:TP330.7。

(4) 复分组配的依据问题。

使用《中图法》进行分类标引时是否结合使用复分、仿分及组配进一步细分,一般应根据类表中的有关注释来决定。其中,只有总论复分表的使用比较灵活,一般可根据需要使用到主表前三级类目或只使用其一部分。对于主表已设有专类或可使用专类复分表、仿分的类目,则不得使用总论复分表。例如,“社会科学论文集”应标引为:C53,不应标引为:C-53;“现代汉语词典”应标引为:H16,不应标引为:H1-61。

(5) 复分的注释方式问题。

类表对仿分或专类复分表的使用,通常有两种注释方式:一种是直接在需仿分的具体类下注明,如在“J617.5 地方戏曲音乐”类下注明“仿 I236 分,必要时再仿 J617.1 分”;另一种是在需要进行仿分的一组类目前做总的注释,如在“G 文化、科学、教育、体育”大类中的“G82/89 各项体育运动”前注明“均可依下表分”,下表是一个专类复分表,可供“G82/89”中的所有类目在必要的情况下使用。前一种注释比较直观,后一种如果不有意识地寻找,有时容易被忽略,因此,在使用后一种注释时,需要对相关的类目体系有比较系统的了解。

5.1.3 分类标引规则

1. 基本分类规则

基本分类规则是整个分类过程中始终必须遵循的具有指导作用的规则,它是从信息资源分类原则中引申出来,并结合分类标引的基本特点和要求确定的。

基本分类规则包括以下几个方面的内容。

(1) 根据信息资源的性质、特点进行分类标引。

对科学知识性质的信息资源标引时,一般来说,主要应根据其内容属性进行,但同时也要兼顾其他特征如国别、时代、形式、类型等。

对文学、艺术形式的信息资源标引时,通常应按照其体裁、形式等进行。

对特定类型的信息资源如工具书、目录、索引、文摘等进行标引时,需要兼顾类目体系的具体规定、用户使用需要、信息资源的内容和形式等方面。

(2) 分类标引时必须能够体现出分类法的逻辑性、等级性、次第性。

凡是能归入某一类的文献,必然要符合其上位类的属性,而不具备这一特点的资源就不能归入该类目。这是由体系分类法从属关系类目的性质决定的。

(3) 必须将特定的信息资源归入最切合其内容的类。

要依据信息资源的内容特征,将其归入分类体系中内涵与外延相一致的、最确切的类目。这是分类法准确、专指地揭示信息资源主题内容的需要。为此,分类标引时既要准确判定信息资源的学科归属,又要按照学科展开的层次,区分总论和专论、理论和应用,以便将其归入最适合的类目。标引人员必须在了解类目体系展开的引用次序及排列次序的基础上,掌握复合主题的设类规律。

(4) 必须将特定的信息资源归入用途最大的类。

满足用户需求是分类标引的最终目的。有些信息资源包含多方面的内容,可以分别归入不同的门类。而不同文献单位由于其专业性质不同,用户的使用要求或习惯不同,对文献处理的要求也存在着差异。分类标引者既要了解文献的内容性质,又要结合文献单位的特点和用户的需求,将文献归入最适用的类目。

(5) 不能单凭题名、篇名的意义归类。

虽然有很多文献的题名可以在一定程度上反映其内容,但也有不少例外,尤其是文艺、社会科学领域的文献,其题名常常只有象征意义,并不能确切反映文献的内容,因此,不能作为分类标引的依据。

(6) 应适当体现分类标引的思想性。

对于社会科学领域的一些具有特殊意义的信息资源,可以有选择地对其内容性质进行适当的揭示。尤其是在网络信息资源的组织中,应适当体现出检索系统的导向性。

2. 一般分类规则

一般分类规则是指从信息资源著作方式的角度提出来的,适用于各个知识门类的分类规则。

信息资源分类的方法与信息资源主题类型、写出版方式等具有密切联系。不同主题类型、写作方式、编辑出版形式等的资源,往往具有不同的标引要求和规律。根据我国目前信息资源分类标引的实际情况,大致可以把一般分类规则概括为以下几个方面。

(1) 单主题信息资源的分类标引。

所谓单主题信息资源,是指论述某一特定事物对象的资源。根据其论述的特点,又可以划分为简单单主题和方面单主题等类型。一般应根据该资源对事物、对象研究的学科角度,按照论述的内容范围进行分类标引。

简单主题信息资源,指只论述一个基本主题对象的资源,一般应按照主题对象的学科性质归类。例如:

“社会利益”,标引结果是:C914。

方面主题信息资源,指论述一个主题一个或多个方面的资源,应根据信息资源论述的方面及各个方面之间的关系归类。例如:

“苹果汁加工方法”,标引结果是:TS255.44。

“苹果树移植”,标引结果是:S661.104。

论述一个主题两个或两个以上方面的信息资源,应根据不同方面之间的关系确定其归属。如果论述的不同方面属于分类法中同一个类别,通常归入其共同的上位类;如果不属于同一个类别,一般应根据信息资源论述的内容重点,归入相应的主要类目;必要时还可以适当进行附加分类。从多个方面全面论述一主题对象的资源,原则上仍应按其主题对象归类。例如:

“电子手表的指示装置与传动机构”,被仿类目中的两个方面的内容属于共同的上位类,应归入其上位类,标引结果是:TH714.52303。

“小麦的生物学原理与营养价值”,两方面内容的类目不属于同一个上位类,不同专业部门可根据需要选择其中一个方面作为主类号,同时为另一个方面作附加分类,两个类号分别是:S512.101;R151.3。

(2) 多主题信息资源的分类标引。

多主题信息资源是指同时论述两个或两个以上的事物对象的资源,一般应按照所论述的主题对象及其关系,区别情况进行分类。根据不同主题之间的关系,多主题资源包括并列关系、从属关系、联结关系主题等基本类型。

并列关系主题的信息资源,是指一信息资源同时论述两个或两个以上各自独立的主题,大致可以分为论及两个主题和论及三个及以上主题两种情况。

论及两个并列关系主题的信息资源,如果同属于类目体系中的同一个类别,就具有共同的直接上位类,通常可直接归入该上位类,否则,可按信息资源的重点或在前的主题归类,同时为另一个主题作附加分类。例如:

“红薯土豆的初加工”,两个并列主题同属于一个类别,可将其归入共同上位类,标引结果是:S530.92。

“局部麻醉与美容术”,两个并列主题不在一个类别之内,可按在前主题归入局部麻醉类,标引结果是:R614.3;同时在美容术类下作附加分类,互见类号是:R622。

将同属于一个类别的并列主题信息资源归入共同的上位类,符合信息资源排架的特点,有利于开架浏览,适合将标引结果同时用于信息资源排架和检索的单位。对于分类结果只供检索的系统,也可以统一按后一种方式标引,这样处理有利于确切检索。

对同时涉及三个或三个以上并列主题的信息资源,一般可根据其涉及的范围,将其归入共同的上位类或概括性类目。例如:

“大白菜、花椰菜、芹菜的施肥技术”,可归入共同的上位类,标引结果是:S630.6。

从属关系主题的信息资源,是指一信息资源同时论述一个大主题和一个小主题,其大主题的外延可以包含小主题。从属关系主题的信息资源一般应按照大主题归类。但如果该信息资源研究的重点是小主题,并没有对大主题展开论述,则可以按照小主题归类。例如:

“雕塑技法与宗教雕塑”,同时论述一个大主题及其从属的一个小主题,应按照大主题雕塑技法归类,标引结果是:J31。

“植物保护与植物检疫”,同时涉及一大主题和一小主题,如果该信息资源的论述重点集中在小主题上,大主题只是作为介绍小主题的背景,则应按照小主题植物检疫归类,标引结果是:S41。

联结关系主题的信息资源,是指一信息资源涉及两个或多个具有联结关系的主题对象,主要包括应用、比较、影响、因果等关系类型。联结关系的信息资源,一般应在分析其关系类型的基础上,按照各自的特点进行标引。

应用关系主题的信息资源,一般应按照被应用到的主题归类。但综合阐述一种理论方法在各个方面应用的信息资源,则应按照该理论方法所在的学科归类。例如:

“太阳能在农业中的应用”,应按其应用领域归入农业科学下的相应类目,标引结果是:S214。

“概率论在经济、军事、历史等领域的应用”,属于综合阐述概率论在各个领域的应用,应按照该理论所在的学科归类,标引结果是:O211.9。

比较关系主题的信息资源一般应按照信息资源作者重点论述的内容归类。例如:

“中美两国保险业的比较”,重点阐述的是中国保险业,应按重点内容归入中国保险业类目,标引结果是:F842。

影响关系主题的信息资源一般应按照被影响的主题对象归类。例如:

“气候变迁对我国农业的影响”,应按被影响的主题归入我国农业类目,标引结果是:S162.22。

因果关系主题的信息资源一般应按照表示结果的主题对象归类。例如:

“染色体异常引起的疾病”,应按表示结果的主题归入染色体疾病类目,标引结果是:R596.1。

(3) 丛书、多卷书的分类标引。

丛书是指汇集多种独立的著作为一套,并具有一个总书名的出版物类型。一般来说,整套丛书或者围绕一个中心问题,或者针对特定的读者对象、用途等编纂。整套丛书中的各个分册的外形基本一致,但内容又往往各自独立,并无多大连贯性。

对丛书进行标引,一般应与其著录方式一致,主要采用两种处理方法:

一种是按集中方式处理,即首先按整套丛书的内容标引,然后再分别按每一个分册作分析标引;

另一种是按分散方式处理,即首先按丛书中的各个分册的内容归类,然后再根据具体情况确定是否为整套丛书编制综合分类款目。

一般情况下,集中归类适合于出版时一次性刊行或虽非一次性刊行,但有明确的出版计划,并且连续出版的丛书。在按整套丛书标引时,除了类表中已经设置相应的丛书专类以外,应在分类号中加上丛书复分号“-51”;如果按照单个分册进行标引,则不加丛书复分号。例如:

《新人文对话录丛书》,集中标引的结果是:C51。

《21世纪经济学丛书》,集中标引的结果是:F0-51,其中分册《财政学》分散标引的结果是:F810。

对于在内容上没有密切联系或者没有明确的出版计划的丛书,应进行分散处理,按照各个分册的内容归类,标引时不加丛书复分号,最后再根据收藏情况及使用的需要确定是否作综合标引。综合标引时,可作一丛书总汇片,此时应加丛书复分号。

多卷书是指将同一著作按若干卷(册)出版的一种文献类型。多卷书通常有总书名,各卷(册)自成一个单位,有的卷(册)有单独书名,有的则没有单独书名。整套多卷书内容连贯,整体性强。所以,多卷书一般应采用按整套书归类的方法。根据多卷书内容组织的情况及出版的特点,通常采用以下两种标引方法。

第一种,对按专题编辑并有分卷(册)书名、各卷(册)具有独立的研究对象的多卷书,一般应在对整套书进行综合标引的同时,以卷、(册)为单位进行分析标引。综合标引时,应加多卷书复分号“-51”。例如:

《中国近代启蒙思潮, 上卷, 启蒙运动的发端: 1840—1914》, 对《中国近代启蒙思潮》进行综合标引的结果是: K25-51, 对《启蒙运动的发端》进行分散标引的结果是: K251。

第二种, 对于无分卷(册)书名、或有分卷(册)书名但各卷(册)并无独立研究对象的多卷书, 只进行综合标引, 标引时可根据具体情况斟酌使用多卷书复分号“-51”。例如:

《从长安到雅典——中外美术考古游记》(上、中、下册), 标引结果是: K869。

(4) 词典、百科全书、年鉴、手册的分类标引。

词典、百科全书、年鉴、手册是汇集一定对象、范围的知识、资源, 按一定方式编排, 供查阅使用的工具书。按照它们所涉及的内容范围, 又可以划分为综合性、专科性两种基本类型。此外, 词典中还包括各种语文词典。这类文献一般应根据其内容范围、出版形式, 并结合文献组织的要求进行分类标引。

综合性词典、百科全书、年鉴、手册等的分类标引, 通常应集中归入“综合性图书”大类, 再按资源类型分入有关门类。例如:

《辞海》, 标引结果是: Z32。

《中国大百科全书》, 标引结果是: Z227。

《中国百科年鉴》, 标引结果是: Z52。

《人民手册》, 标引结果是: Z52。

专科性词典、百科全书、年鉴、手册的分类标引, 有两种基本处理方法: 一种是分别按内容归入有关的知识门类, 再使用总论复分号揭示其资源类型; 另一种是将其集中于“综合性图书”大类的相应专科类目下, 再按照类表的规定以组配方式揭示其学科内容。一般情况下, 在集中建立工具书典藏的文献单位, 应采用集中处理的方法; 否则, 采用分散处理的方法。例如:

《中国出版年鉴》, 分散标引的结果是: G239.2-54, 集中标引的结果是: Z58: G239.2。

语文词典的分类标引, 一般应根据其特点归入语言类, 按该类的相应规定分。

按照《中图法》的规定, 一种语文的词典, 应归入该语文的词典、辞典类; 两种外语对照的语文词典, 归入在前的亦即被解释的语文, 同时在后一种语文下作附加分类; 汉语与少数民族语言或外语对照的词典, 不管在先的是否是汉语, 均归入相应的少数民族语言或外语; 三种以上语言对照的语言词典归入语言文字类的“H061 词典”类; 但专科词典, 不管它包括多少种语言, 一律应按专业内容归入有关学科门类。例如:

《现代汉语词典》, 标引结果是: H164。

《俄日词典》, 标引结果是: H356, 附加分类标引结果是: H366。

《汉英双解词典》, 标引结果是: H316。

《蒙汉词典》, 标引结果是: H212.6。

《汉语成语词典》, 标引结果是: H136.3-61。

《英法汉会话词典》, 标引结果是: H061。

《英汉对照经济学词典》, 分散标引的结果是: F0-61, 集中标引的结果是: Z38: F0。

(5) 目录、索引、文摘的分类标引。

目录、索引、文摘是提供文献查找线索, 指导阅读的工具书。根据其揭示特点, 这类资源可以分为综合性、专科性、专书或专题等类型, 通常应根据其揭示对象及范围, 按照文献组织的需要分类标引。

综合性目录、文摘、索引的分类, 通常应归入“综合性图书”大类的相应门类。例如:

《全国总书目》, 标引结果是: Z812.1。

《全国中文期刊联合目录》, 标引结果是: Z87。

专科性目录、文摘、索引的分类, 通常采用两种做法: 一种是将其集中于“综合性图书”大类的有关门类, 再按组配号法揭示其学科; 另一种是根据本单位文献组织的特点, 按照

其内容分散归入有关各类,并在分类号后再加总论复分号“-7”。例如:

《中国古代雕塑考古著作目录》,集中标引的结果是:Z88:K879.3,分散标引的结果是:K879.3-7。

《敦煌学论文索引》,集中标引的结果是:Z89:K870.6,分散标引的结果是:K870.6-7。

《医学文摘》,集中标引的结果是:Z89:R,分散标引的结果是:R-7。

专书索引一般应随原书归类,以方便用户将其与原书结合使用。对于马列经典作家的著作及研究的书目、索引,《中图法》已在马列大类设有专类,应归入相应专类。例如:

《资治通鉴人名索引》,标引结果是:K204.3。

《鲁迅文学作品书目》,标引结果是:I210.99。

《毛泽东选集目录》,标引结果是:A843。

(6) 关于对著作的研究、注释文献的标引。

该类文献包括对特定著作的评论、研究、注释、翻译、校勘、考证、改编等多种类型,一般应结合该类文献的形式特点及分类法中类目设置的情况处理。

科学著作的评论、研究、注释,通常按内容与原书归入一类,必要时,再使用专类复分表区分其著作方式;但为满足语言学习的需要而编制的文献,分类时应考虑其使用目的,归入语言大类中的相应类目。例如:

《本草纲目注释》,标引结果是:R281.3。

《应用植物学(英汉对照)》,标引结果是:H319.4或H319.4:Q949.9。

马列经典作家著作的评论、研究,应按照《中图法》的类目设置,归入马列大类下有关评论、研究的专类,但运用马列主义的思想研究和解决各领域问题的文献,应按讨论的问题归入相应的知识门类。例如:

《学习〈湖南农民运动考察报告〉的体会》,标引结果是:A841.22。

《列宁图书馆理论的现实意义》,标引结果是:G250。

文学作品的评论、研究,应根据文学类的设类特点,按其研究对象的国别、体裁归入各体文学评论研究的相应类目。例如:

《红楼梦研究综述》,标引结果是:I207.411。

缩写、节选的文献,如果内容性质未发生变化,仍按原书归类,但如果内容性质已发生较大变化,则应重新归类;改编的文献,如果将一种体裁的文艺作品改写为另一种体裁的文艺作品,一般应按照改写后的体裁归类。例如:

根据王安忆长篇小说改编的同名电视连续剧剧本,应归入建国后电视剧,标引结果是:I235.2。

(7) 特种文献的分类标引。

技术标准是各国政府部门或国家行业组织等对工农业生产及工程建设质量规格等所作的技术规定。按照使用范围,技术标准一般可分为国际标准、专业标准、企业标准;我国现行的技术标准主要分为国家标准、部颁标准。

专利文献是指经专利机构批准的,获得专利权的新技术、新方法、新工艺、新产品的文献形式。专利文献及时报道各种科学技术发明的内容,是科技工作者和科研人员了解世界各国的科研水平的重要信息资源。

各个文献单位对于技术标准、专利文献的分类标引的方法不尽相同。专门文献单位往往使用专门类表如《国际专利分类表》等作为分类标引工具,一般文献单位则往往按照通用的分类法内的有关规定进行处理。例如:

《文献著录总则》(GB 3792.1—83),采用《中图法》对其进行分散标引的结果是:G254.31,集中标引的结果是:T-652.1。

《焊接技术专利》，采用《中图法》进行标引的结果是：TG4-18。

科技报告是有关一专题研究进展或科学技术成果的报告。学位论文是高等学校或科研单位学生在申请学位时提供的论文。无论是科技报告还是学位论文，对于科学研究都具有较重要的参考价值。文献单位对这两类信息资源一般采用单独设库的方式收藏管理。

科技报告和学位论文的分类与其他类型信息资源相同，一般按论述对象的学科角度归类。其中，因为学位论文的内容往往比较专指，所以，如果数量较多，可以直接使用《中国资料分类法》进行标引，如果数量不多，则直接使用《中图法》。例如：

《世界各国环境综合调查研究》（科技报告），标引结果是：X508。

《图书馆读者心理学综论》（硕士论文），标引结果是：G252。

（8）非书资料的分类标引。

非书资料又称为非印刷型资料、非纸质资料，可以细分为缩微型、视听型、机读型、光盘型等多种类型。非书资料借助现代信息技术使得信息媒体在记录方式、记录对象、使用功能等方面产生了质的飞跃。而且，随着现代信息技术的发展，文献单位收藏管理的非书资料的数量呈迅速上升的趋势。

非书资料作为一种与印刷型文献不同的媒介形式，在文献单位中通常按其形式与传统文献分别保管，不同于印刷型文献多采用分类方式组织的特点，对非书资料基本采用固定排架方式组织。但是，为了满足用户按照内容检索的需要，同样应以非书资料的内容学科属性为主要依据，编制分类检索系统。所以，在使用分类法主表的基础之上，再依据总论复分表揭示其媒介形式，是当前对非书资料进行分类标引的主要方法之一。例如：

《失落的文明——古希腊古罗马历史》（激光视盘），使用《中图法》进行标引的结果是：K12-793。

（9）网络信息资源的分类标引。

网络资源的特点是数量大、种类多、动态性强、内容分布范围广。网络信息资源既包括正式出版物，又包括大量灰色文献、个人信息；既包括已有的传统信息资源类型，还涉及BBS、聊天室、新闻组、多媒体资源等多种形式。网络信息资源在内容的分布上，涉及新兴科学技术、商业、娱乐等的相对比较较多。

目前，对于网络信息资源的分类，通常采用两种方式：一种是在采用传统分类体系的基础上，进行必要的增补，例如，国外依据DDC、UDC、LCC等建立的一些网络分类检索系统，采用的就是这种方式；二是直接使用以网络信息资源为对象编制的分类体系，例如，Yahoo!、搜狐、蓝帆等网络分类检索系统，采用的就是这种方式。

各种网络信息资源的分类方法与传统信息资源处理的方法基本相同，一般都是根据分类体系的特点，将特定信息资源归入与其相对应的类目之下。但是，在第一种方式中，可以将传统文献分类法的类目体系作为基础，根据网络信息资源的具体情况，对原有的类目体系进行有限的调整，以尽可能地适应网络信息资源的分类需要。例如：

南京大学信息管理系网站，使用《中图法》对其进行标引的结果是：G23-40（2），G25-40（2），G27-40（2），G35-40（2）。

5.2 主题标引

5.2.1 主题标引概述

1. 主题标引的概念

主题标引是依据一定的主题词表或主题标引规则，赋予信息资源语词标识的过程。具体

而言,主题标引是在主题分析的基础上,以一定的主题词表或主题标引规则作为依据,将信息资源中具有检索意义的特征转化成相应的主题词,并将其组织成表达信息资源内容特征的标识的过程。按照是否使用主题词表,主题标引分为受控标引和自由标引两类。

2. 主题标引方式

主题标引方式是根据信息资源的特点和使用需要确定的标引和揭示信息资源主题的形式。主题标引方式一般分为以下几种类型。

1) 整体标引

整体标引(亦称浅标引)是一种概括信息资源基本主题内容的标引。整体标引的对象可以是图书、论文、标准、档案或其他信息资源类型。这种标引只揭示信息资源中具有检索价值的整体性主题,不揭示其涉及的各种从属性主题内容。例如,对《信息组织概论》一书进行整体标引时,只需对“信息组织”这一主题内容进行标引即可。如果某一信息资源同时涉及两个或两个以上整体主题,则应根据情况,按照多主题的标引要求进行标引。

例:对于《分类法与主题法》这一文献,因其涉及到两个整体主题,可标引为:

分类法

主题法

2) 全面标引

全面标引(亦称深标引)是一种充分揭示信息资源所论及的所有有检索价值的主题概念的标引。这种标引不仅要求揭示信息资源论述的整体主题,而且要求揭示符合检索系统要求的所有主题概念。例如,《文献自动分类因果推理技术的实现》一文,结合《中图法》和《中国分类主题词表》,分析其内容、结构,形成字段分类因果推理规则,实现文献的自动分类,从而解决了目前自动分类的瓶颈问题。在对其进行全面标引时,就必须将:自动分类、因果推理机、语义层次结构、中图法、中国分类主题词表等主题概念一一加以标引和揭示。

全面标引可加深对信息资源内容的揭示程度,有利于提高查全率,但需花费较大的人力物力,适用于主题标识结合机检系统处理专业领域的论文、技术报告等类型文献。以图书为对象的标引一般不采用全面标引。在实际操作中,主题词的标引数量通常保持在5~20个。

3) 对口标引

对口标引(亦称重点标引)是一种只揭示信息资源中适合本专业需要的主题内容的标引。例如,“泰国的电子行业和汽车行业”这一主题,电子行业的研究单位可以用电子工业、泰国两个主题词进行对口标引;汽车生产研究单位可以用汽车行业、泰国两个主题词进行对口标引。对口标引具有较强的针对性,可使标引工作更好地适合用户的实际需要,提高查准率。有明确服务对象的专业文献单位常采用这一标引方式。

4) 综合标引

综合标引是一种以丛书、多卷书、论文集、会议录、标准汇编、档案的案卷等为单位进行的概括性标引。综合标引除揭示信息资源的主题内容外,一般还应根据情况对信息资源的类型进行必要的揭示。例如,《计算机科学丛书》在以整套书为单位进行标引时,除对整体内容“计算机科学”标引外,通常还对“丛书”这一类型进行必要的揭示,应标引为:

计算机科学—丛书

5) 分析标引

分析标引是一种根据信息资源中部分片段或集合型信息资源的构成单位进行的标引。例如,《网络信息优化传播导论》一书,除了对整体主题“网络传播”进行标引外,还可将其其中单元的“网络媒体评价”内容析出进行分析标引。分析标引是与整体标引和综合标引相对应的标引方式,其作用在于可以在进行整体标引或综合标引的同时揭示信息资源中有检索价

值的主题内容。

采用何种方式进行主题标引，通常应结合检索系统的设备条件、信息资源的特点、收藏范围、用户需要等多种因素加以综合考虑。例如，手工检索工具采用先组式标引，对信息资源内容的揭示比较概括，宜采用整体标引，必要时还可进行分析标引。机检系统以后组式标引为主，通常对主题的揭示比较充分和完备，宜采用全面标引。

3. 主题标引程序

1) 查找利用已有的标引成果

即要查明被标引的信息资源是否已被标引过，有无标引成果可以直接采用或作为参考。目前可利用的标引成果很多，主要有：本单位的标引成果（查看标引文献是否为本单位收藏并标引过的文献的复本）；再版标引成果（再版编目数据中的主题词）；外单位的标引成果（以联机方式在特定网络中查找本单位没有标引但其他单位已经标引了文献的主题标识供本单位利用）。利用已有的标引成果，可以保证信息资源标引的一致性，减少标引误差，提高标引速度。

2) 主题分析

如果没有现成的标引成果可利用，则要进行标引。为此，需要对信息资源进行主题分析。基于人工的主题标引，要在充分了解信息资源内容及研究对象的基础上，对主题的类型、结构及其构成要素等进行深入分析，对有检索意义的主题概念进行概括、提炼和选择。基于自动标引的主题分析则表现为在信息资源中抽取表达主题的自然语词的方法的运用，如词频统计分析、语词位置加权等。

3) 主题概念的转换

用标引语言（标题语言、叙词语言等）的标识（标题词、叙词等）表达主题概念的过程称为主题概念的转换。人工标引中主题概念转换首先要辨识标引工具（标题表、叙词表）中的相应标识（标题词、叙词）的含义，然后选择恰当的标识。自动标引中的赋词标引，其主题概念转换是由计算机将文献中能表达主题的词与主题词进行相符性比较而完成的，将自然语言转换为主题语言。

4) 主题标引记录

主题标引记录就是在规定的载体上，按照一定的格式，对主题标引的结果和标引中所处理的一些重要问题做出记录。一般而言，在卡片或书本式索引中，主题标引的结果记录在卡片或书本式目录排检位置，在计算机检索系统中，则记录在文档的相应字段。

5) 审核

主题标引结果的审核是主题标引的最后一道工序，也是保证主题标引质量的一个重要环节。审核包括以下几个方面的内容：信息资源主题的提炼是否全面、准确，有无遗漏隐含的主题和有潜在用途的主题；标引方式是否符合检索系统和信息资源类型的要求；选用的主题词是否符合标引规则；等等。

5.2.2 主题标引方法

1. 主题分析方法

主题分析就是根据信息资源存储与检索系统的需要，对信息资源的内容进行分析和提取主题概念的过程。其目的是在了解、判断信息资源内容的基础上，确定信息资源的主题，确定各主题之间及构成主题的因素之间的关系，并从概念上加以提炼，以便据此选择一组恰当的、语义相应的主题词作为信息资源的主题标识。主题分析是主题标引的重要基础，主题标

引质量的好坏、检索效率的高低,首先取决于主题分析的优劣。要准确、恰当地进行主题分析,首先,需要了解和掌握信息资源的主题类型;其次,需要掌握主题的结构;再次,需要掌握主题分析方法。

(1) 主题类型的分析。

主题是信息资源论述的中心问题,是概括信息资源中心论题的一个概念或若干个概念的组,根据不同的标准,可以划分出多组主题类型。

① 单主题和多主题。

这是根据某一信息资源中讨论的主题数量划分的主题类型。单主题指信息资源中只研究一个中心对象或问题,它可以概括论述某一特定的事物或问题,也可以论述事物对象的某一部分或某一方面问题。例如,互联网络、环境保护、信息咨询、互联网络的安全、环境保护的政策、信息咨询的方法等。对于单主题还应进一步根据其主题构成分析其成分。多主题指某一信息资源同时研究两个或多个事物对象或问题,它们之间可以是并列关系,如传统图书馆和数字图书馆;也可以是同时论述一个大主题和若干个小主题的从属关系,如信息检索与社科信息检索、科技信息检索。

② 单元主题、复合主题和联结主题。

这是根据主题概念的数量及其关系划分的类型。单元主题指主题中只含有一个基本概念构成的主题类型,如教育学、数据挖掘等,对单元主题的分析,只要用一个概念加以概括就可以了。

复合主题指主题中含有两个或两个以上的主题因素,也称多元主题。复合主题是比较复杂的主题,根据主题因素之间的关系,又可分为并列主题概念组配构成的复合主题、事物与其方面概念构成的复合主题、事物与其部分概念构成的复合主题等不同类型,如宽银幕彩色电影、城市交通的管理、飞机发动机等。对复合主题,要在认真分析其组成的要素与要素之间关系的基础上,再对主题概念进行取舍。

联结主题亦称相关主题,是指两个或多个主题对象之间根据一定的联系所形成的一种主题类型。在这类主题中,不同主题对象之间的关系比较松散,不像一般复合主题那样已经融合成一个整体,是一种介于单主题与多主题之间的主题类型,常见的关系类型有:

- 应用关系,如数学方法在信息管理中的应用;
- 比较关系,如中美文化比较;
- 影响关系,如风速对运动成绩的影响;
- 因果关系,如环境污染会导致气候变化;
- 倾向关系,如供旅游用的日语读本。

③ 主要主题和次要主题。

这是依据主题对信息资源内容概括的重要程度划分的类型。主要主题是信息资源论述的主题内容中作者重点论述的主题。一个信息资源一般至少有一个主要主题,有时也可以有两个或多个主要主题。例如,“系统论、信息论和控制论”一书中,就包括系统论、信息论和控制论三个主要主题。次要主题是指不作为重点论述的主题,对于次要主题可根据检索系统的要求及其本身的情报价值确定是否将其析出。

④ 专业主题和非专业主题。

这是依据主题与检索系统专业的相关程度划分的类型。专业主题是指与检索系统专业性一致的主题,它可以是主要主题,也可以是次要主题,对于这类主题,专业检索系统一般应充分揭示。非专业主题是指与检索系统专业性不一致的主题,对于这类主题,专业检索系统往往不予揭示或只在对该专业的研究有联系、有启发、有一定使用价值时才酌情析出。

⑤ 显性主题与隐性主题。

这是依据主题对信息资源概括的清晰程度划分的类型。显性主题指信息资源正面明确阐述的主题。隐性主题则指信息资源中没有用直接的语词明确描述，隐含在其他字面形式中的主题。例如，“信息组织的系统论基础”这一主题，其显性主题是信息组织、系统论，其隐性主题是分类法、主题法。隐性主题容易被遗漏，应在深入了解信息资源内容、了解有关领域主题内容之间的关系和影响的基础上提炼出来，这类情况在深度标引时出现较多。

（2）主题结构的分析。

主题结构是指构成信息资源主题的各因素及它们之间的相互关系。任何信息资源的主题（除了纯粹的单元主题外）都是由一定的主题因素构成的，主题因素之间存在着一定的结构关系。从主题标引的角度把主题的结构形式加以模式化，有助于准确并比较一致地进行主题标引。

主题结构在国内外有较广泛的研究，已经有了比较成熟的主题结构模式（即引用次序）。国外比较著名的引用次序有：①显著性引用次序：事物—部件—材料—活动—施动者；②范畴职能引用次序，是指将各种主题概念划分为范畴，并按范畴的性质、职能确定组配次序，其中最著名的是印度图书馆学家阮冈纳赞的范畴分面公式和英国分类法研究小组维克利提出的标准引用次序。阮冈纳赞范畴分面公式为本体—物质—动力—空间—时间；维克利的标准引用次序为物质（产品）—种类—部分—成分—性质—过程—操作—施动者或工具。对于中文引用次序，国内主要采用刘湘生的主题分面公式，将信息资源的主题和事物各种属性特征归纳成5个基本方面，每个方面即为一种类型的主题因素，并规定了5种主题因素的代码及引用次序：A 主体因素—B 通用因素—C 位置因素—D 时间因素—E 文献类型因素。对主体因素还可细分，并规定组配次序如下：A1（对象）—A2（方面）—A3（方法）—A4（结果）—A5（条件），这样就把原引用次序进一步展开为：A（A1—A2—A3—A4—A5）—B—C—D—E。在这个公式中，主体因素是指能反映主题中主要特征属性的一组主题概念，这些主题概念都具有独立的检索意义，具体包括：研究对象、方面、方法、结果、条件等因素。通用因素是反映主题中一般通用特征属性的概念因素，只对主体因素起修饰说明作用，无独立检索意义。位置因素是反映信息资源主题中空间地理属性的概念，包括自然区域和行政区划区域等方面的概念因素，一般不具有独立检索意义。时间因素是反映信息资源主题中所处的时间属性的概念。文献类型因素是指表现信息资源类型的概念，如手册、索引、词典等。刘湘生的主题分面公式已被写入国家标准《文献主题标引规则》（GB 3860-83），它是我国主题标引工作中第一个比较完整的引用次序。

2. 主题概念的转换

受控主题标引通常是在主题分析的基础上，依据一定的主题词表，将分析出的主题概念转化为规范化的主题词，并根据检索系统的要求，对标识进行处理。主题概念转换的方式通常分为两种：主题概念的直接转换和主题概念的分解转换。

（1）直接转换。

是指分析出来的主题概念可以直接用主题词表上的一个对应主题词加以表述。如可将“市场营销”这一主题概念直接转换成相应的主题词“市场营销学”。

（2）分解转换。

是指分析出的主题概念在主题词表中没有现成的主题词直接表达时所必须采取的一种方式，即首先将复杂主题概念分解成若干个基本概念，再从主题词表中选取与基本主题概念对应的主题词，并按一定的组配规则组合起来表达复杂的主题概念。进行主题概念分解应注意以下问题。

① 应采用概念分解,避免字面分解。在用叙词语言进行主题标引时,由于叙词语言的基本原理是概念组配,而概念分解是概念组配的逆过程,因此在分解转换时应采用概念分解。如“无机药物化学”应分解为无机化学和药物化学,而不能分解为无机和药物化学。

② 必须采用最专指的概念分解形式。如“农业经济结构”最专指的分解形式是农业经济和经济结构。

③ 应采用交叉关系分解优先的原则。这是指当一个复杂的主题概念既可以采用交叉关系分解方式,又可以采用限定关系分解方式时,依据概念组配的原则,优先采用交叉关系分解方式。如“植物生物化学”应分解为“植物学:生物化学”,而不应分解为“植物—生物化学”。

3. 主题标识的确定

标识的确定指依据检索系统的使用需要,在完成标识转换的同时,对标引词进行必要的处理。目前国内的主题检索工具大致有两种类型:其一是书本式索引或卡片式目录等手检工具;其二是计算机检索系统。手检工具大多采用先组方式,一般应按照检索工具的要求,将主题词组织为标题。机检系统一般采用后组方式,通常要求使用适合后组方式的句法手段进行处理。

(1) 确定标题。

标题是一种以先组方式表达信息资源内容的标识形式,利用叙词语言编制手工检索工具时,一般应根据手工检索工具的需要,对标引词进行处理,确定标题。

① 确定标题的结构形式。

根据主题标引时使用的主题词数量和组配符号的不同,标题一般可分为单级标题和多级标题。单级标题是由一个主题词构成的标题,包括单词标题、词组标题、带限义词的标题等,如信息学、网络安全、运动(哲学)等。多级标题又叫复合标题,是由两个或多个主题词通过一定语义逻辑关系组配形成的,在《中国分类主题词表》第二版中都用“\”作为组配符号,如土壤微生物学\土壤生态学、残疾人\社会问题、经济作物\病虫害\预测等。

② 确定标题的级别。

标题的级别决定着反映信息资源主题的专指程度,直接影响检索工具的查准性能。目前国内手工检索工具大致有3种级别。即只采用单级标题,采用二级标题和采用三级以上标题。采用多级标题形式可以对信息资源进行充分的揭示,比较符合现代信息资源的揭示要求。因此我国手工检索工具一般将标题规定为三级或三级以上,一般不超过五级。在采用二级或多级标题时,还要注意以下几个问题:① 做好主标题的选择,主标题是标题的检索入口,关系到检索系统的排检位置,因此主标题一般应具有独立的检索意义,通常为表示主体因素的叙词;② 对叙词的引用次序做出规定,在我国一般依据刘湘生的主题分面公式确定;③ 规定轮排模式,即依次将标题中每个有独立检索意义的叙词轮流排主标题,标题中其他词的位置保持不变。通过轮排可以提供更多的检索入口。

(2) 机检词的处理。

主题标引的目的,除了为手工检索系统的主题目录、主题索引拟定标题外,另一主要目的就是作为机检系统的主题标识,输入机检文档。叙词具有可组配、灵活检索的特点,而且机检系统中主题标引通常为深标引,采用后组方式,因此为了避免叙词之间可能出现的错误组配,通常对机检词进行以下处理。

① 加联号。联号是表明同一信息资源的多个标引词中,哪些词之间有联系而可以互相组配,哪些词之间没有联系而不应进行组配的符号,一般用数字或字母表示,其作用就是排除主题标识之间虚假组配的可能。例如,“全球的气候变化和亚洲的生态环境”这一并列主

题文献,如果不对标引词进行处理,则有可能会出现“全球”与“生态环境”、“亚洲”与“气候变化”的错误组配,使用联号就可避免上述错误的组配,假设该主题文献的文献号为1235,联号用数字1、2表示,则该文献可标引为:

主题标识	联号	文献号
气候变化	1	1235
全球	1	1235
生态环境	2	1235
亚洲	2	1235

这样,检索比号时,不仅文献号要相同,而且同一主题的关联符号也必须相同,这就不会造成误检。

② 加职号。职号是表示标引词在组配表达主题时的职能作用的符号,其作用是指明同一主题的标引词中,某一词的职能或角色,从而提示词与词之间的关系,明确它们所表达的概念,提高查准率。使用职能符号一般应根据需要预先确定句法范畴和相应的符号,例如:

符号	职能
A	动作对象
B	部分
C	性质
D	操作
E	施动者

例如,对“森林对气候的影响”这一文献标引时,如果不对标引词进行处理,则有可能会出现“森林对气候的影响”和“气候对森林的影响”两种含义,而使用职能符号就可以避免这种现象。假设该文献号是346,应标引为:

森林 E	346
影响 D	346
气候 A	346

(3) 标引词的著录

根据检索工具的不同特点,标引词可分别采用相应的著录格式。手检系统中标引词的著录有两种情况:一种是将标题直接著录在目录卡片上,作为排检依据;另一种是直接将文献号记录在相应标目下,主要用于主题索引。

机检系统的著录应根据系统要求的方式进行。文献单位一般根据使用需要,设计标引工作单作为著录的依据,著录时应严格按照工作单的格式及相应规定进行。在填写标引单的同时,应根据要求做好辅助符号如联号的标注。利用机读文档自动输出目录卡片或编制书本式目录的单位,还应在输入机检词的同时,为准备输出的手检标题标注相应的主、副标题符号。

为了方便主、副标题输入的统一,我国国家标准《文献叙词标引规则》(GB/T 3860—1995)规定,主、副标题的符号可依据英文名称的首字母分别表示如下:

- M(MAIN HEADING)——表示主标题;
- Q(QUALIFIRE)——表示副标题;
- S(SUBQUALIFIRE)——表示三级以上副标题。

在三级以上副标题中,又可以根据需要进一步分出:

- SA——表示三级标题;
- SB——表示四级标题;
- SC——表示五级标题。

根据上述规定,机编系统就可按 M—Q—SA—SB—SC 的次序自动生成复合标题。

5.2.3 主题标引规则

为了客观、正确、全面地揭示信息资源的主题内容,保证主题标引的质量,必须制定具体的标引规则。作为主题标引的依据,国家标准《文献叙词标引规则》(GB/T 3860—1995)详细论述了叙词标引的基本原则和要求。

1. 主题标引的基本规则(以叙词语言为例)

1) 选词规则

(1) 选用词表中的正式叙词进行标引,其书写形式要与词表中的词形完全一致,非叙词不得直接用来标引,只起指向正式叙词的作用。例如:

逻辑代数

Y 布尔代数

一篇关于“逻辑代数研究”的文献,其标引词应用“布尔代数”、“研究”,而不是“逻辑代数”、“研究”。

(2) 选用与主题概念相对应、最专指的叙词标引。如一篇论述“超声学”的文献,既不能用其上位词“声学”标引,也不能用其下位词“低温声学”标引,而只能用专指叙词“超声学”进行标引。

(3) 组配标引。当词表中没有相应专指叙词时,可选用词表中最接近、最直接关联的两个或两个以上的叙词进行组配标引。如“石英电子手表”,词表中没有这个专指叙词,应选用“石英手表”和“电子手表”两个最直接相关的叙词进行组配标引。

(4) 上位词标引。当词表中没有最专指的叙词,也无法进行组配标引时,可选用一个最接近的上位词进行标引。如“柜类家具”这个主题概念,词表中既无专指叙词,又不适合用“柜类”和“家具”组配,可直接选用“家具”这个最接近的上位叙词进行上位词标引。

(5) 靠词标引。即使用含义相近的叙词进行标引。当主题概念在词表中既无专指叙词,又不能进行组配标引、上位词标引时,可进行靠词标引。如“图书注销”这一主题概念,词表中既无专指叙词,又不适合组配标引、上位词标引,可采用与它含义最接近的相关叙词“图书登记”进行靠词标引。

(6) 增词标引。当主题概念采用专指词标引、组配标引、上位词标引、靠词标引等标引方式均不合适时,可以考虑增补叙词标引。新增的叙词必须是词表中明显漏收的重要主题概念,或是表达新科学、新理论、新技术、新材料、新发展的词,或者是组配标引中可能出现二义性的词,而且新增词必须词形规范、概念明确、具有较强的组配作用。

(7) 自由词标引。自由词一般是指不受词表控制,用于自然语言标引和检索的自然语言词汇,自由词标引是一种重要的标引技术,它可以作为主题标引的补充和辅助手段,弥补主题标引的不足。自由词标引主要用在计算机检索系统中,有自动标引和人工标引两种方式。自由词的自动标引主要是从题名和文摘中自动抽取一定频率或权值的词来揭示信息资源的主题内容。自由词的人工标引,一般由标引人员抽取出现在信息资源中的、词表中遗漏的主题概念,或者是词表中未收录的专有名词和细小的专指词等。选用的自由词,应当做到词形简练、概念明确。同时,要做好相应的记录,以保证标引用词的一致性。在手工检索系统中,为方便读者查询,自由词标引还可增做一张Y项参照指导片,分别排在相应的字顺位置。在计算机检索系统中,则应将标引用的自由词著录于610字段。

对于一个特定的主题来说,一般只选择增词标引或自由词标引中的一种,对于一个特定的检索系统而言,可同时允许增词标引和自由词标引。无论是增词标引还是自由词标引,均应遵守相应的具体规则。

2) 组配规则

组配标引是将两个或两个以上叙词按照一定的逻辑关系结合在一起,表达主题概念。为了保证在组配时尽可能一致,一般应遵循下述组配规则。

(1) 叙词组配必须是概念组配。即要求参加组配的叙词之间应具有概念交叉、概念限定等逻辑关系,而不能采用简单的字面拼合和随意组配。如“肝外科手术”应标引为“肝疾病”和“外科手术”不能标引为“肝”和“外科手术”。

(2) 叙词组配应优先采用交叉组配。只有在不能用交叉组配时,才用限定组配。如“介质光波导”应采用“介质波导”和“光波导”进行交叉组配,而不能用“介质”和“光波导”进行限定组配。

(3) 不能越级组配。即在可以用相应专指叙词组配时,不得使用该词的上位叙词或下位叙词进行越级组配。如“道路运输经营学”可用“公路运输”和“运输经济”组配标引,而不能用“交通运输”和“运输经济”或“公路运输”和“经营管理”组配,“交通运输”和“经营管理”是上位词,属于越级。

(4) 叙词组配必须概念清楚、确切,具有单义性。如果组配的结果产生二义现象,可根据需要采用增词标引或增加词组中介叙词等方法解决,如“知识经济”这一主题不能用“知识”和“经济”组配标引,因为可能产生二义性,即知识经济或者是关于经济的知识,可通过增补“知识经济”这一新词的办法解决。又如“体育学校”也不能用“体育”和“学校”组配,因为也可能有两种含义,即体育学校或者学校的体育,为防止歧义现象,可通过增加中介叙词的办法解决,如“体育—教育机构—学校”。

(5) 叙词组配次序。应按主体因素—通用因素—时间因素—地区因素—文献类型因素的次序确定。如“我国 20 世纪 90 年代汽车工业规划研究文集”,按上述次序确定为:汽车工业—规划—中国—1990—1999—文集。

2. 主题标引实例

以下以 CALIS 联合目录中文文献书目数据中的主题标引为例,以《汉语主题词表》为标引依据,采用 CNMARC 机读格式。

(1) 《全国高等学校图书馆工作会议文集》标引为:

606 0# \$a 院校图书馆 \$x 图书馆工作 \$y 中国 \$j 文集

注:① 在《汉语主题词表》中,大学图书馆 Y 院校图书馆

因此应选用词表中的正式叙词“院校图书馆”进行标引。

② \$a 表示主体因素,\$x 表示通用因素(包括论题复分),\$y 表示位置因素,\$z 表示时间因素,\$j 表示文献类型因素。

(2) 《心电图诊断技巧》标引为:

606 0# \$a 心电图\$x 诊断

不能标引为:606 0# \$a 电诊断。

注:在《汉语主题词表》中,电诊断

F 心电图

因此应选用词表中最专指的叙词进行标引。

(3) 《计算机在企业管理中的应用》标引为:

606 0# \$a 企业管理 \$x 计算机应用

不能标引为:606 0# \$a 企业管理\$x 电子计算机\$x 应用。

注:应选用词表中的复合主题词“计算机应用”进行标引。

(4) 《中等师范教育》标引为:

606 0# \$a 师范教育\$x 中等专业教育

注：该例为组配标引，其组配成分（交叉组配、限定组配）都记录在子字段\$x。

(5)《石英电子钟表修理大全》标引为：

606 0# \$a 石英钟 \$x 电子钟 \$x 维修 \$j 手册

606 0# \$a 石英表 \$x 电子表 \$x 维修 \$j 手册

注：该例为两个并列主题的组配标引，其组配成分（交叉组配、限定组配）都记录在子字段\$x。

(6)《香蕉苹果的栽培》标引为：

606 0# \$a 苹果 \$x 栽培

610 0# \$a 香蕉苹果

注：该例为上位词标引，并同时进行自由词标引，记录在 610 非控主题词字段。

(7)《图书倒架问题》标引为：

606 0# \$a 图书排架

610 0# \$a 图书倒架

注：该例为靠词标引，并同时进行自由词标引，记录在 610 非控主题词字段。

(8)《视听新潮流：家庭影院》标引为：

610 0# \$a 家庭影院

注：该例为增词标引，新增词（待正式批准）暂记在 610 字段。

(9)《概率论在经济、军事、地理领域的应用》标引为：

606 0# \$a 概率论 \$x 经济 \$x 应用

606 0# \$a 概率论 \$x 军事 \$x 应用

606 0# \$a 概率论 \$x 地理 \$x 应用

注：该例为多主题文献的标引，一般将多主题分解为单主题，进行分组标引或分组合配标引。一般不选用上位词标引，但如果标引深度超过本单位的规定时，可重点选择其中几个并列的单主题进行标引，同时再标引一个它们的上位主题。

(10)《英国概况》标引为：

607 ## \$a 英国 \$x 概况

注：当位置因素已成为文献的主要研究对象时，应确定为主标题，记录在 607 地理名称主题字段。

5.2.4 主题标引与分类标引的比较

标引是将信息资源主题的自然语言转换成规范化的检索语言的过程，也就是对信息资源进行主题分析并赋予检索标识（即分类号、主题词）的过程。主题标引和分类标引是信息资源标引的两种基本方法，它们都是从语义角度对信息资源进行揭示和组织的。主题标引是依据一定的主题词表或主题标引规则，赋予信息资源语词标识的过程。具体而言，主题标引是在主题分析的基础上，以一定的主题词表或标引规则为依据，将信息资源中具有检索意义的特征转化成相应的主题词，并将其组织成表达信息资源内容特征的标识的过程。按照是否使用主题词表，主题标引分为受控标引和自由标引两类。分类标引则是依据一定的分类语言如分类表，对信息资源的内容特征进行分析、判断、选择，并赋予其分类标识的过程。因此主题标引和分类标引是两种不同体系结构的标引方法，两种方法既存在重要差异，也有一些共同之处。下面对主题标引和分类标引之间的异同进行分析比较。

(1) 主题标引与分类标引的标引对象相同，但揭示信息资源内容的角度不同。

主题标引和分类标引的标引对象是相同的，都以信息资源的主题内容为揭示和转换对象。无论是主题标引还是分类标引，都必须首先对信息资源进行主题分析，形成主题概念，然后将主题概念转换为主题词或分类号。只不过主题标引是按主题法从特定事物方面来揭示和组织信息资源的，而不管其在科学体系中的位置。分类标引是按分类法从信息资源内容的学科属性来系统地揭示和组织信息资源的，并从知识分类的角度揭示各类信息资源在内容上的区别和联系，这是主题标引与分类标引的主要差异。

(2) 主题标引与分类标引在标引时所使用的标识符号不同，使得主题标引具有直观性，分类标引具有间接性。

主题标引采用直接的语词标识系统，以规范化或不经规范的自然语言为信息资源内容主题的标识符号，这种标识符号比较直观，使人一目了然，使得主题标引具有直观性。分类标引采用间接的号码标识系统，即以字母、数字或两者混合的号码为标识符号，在标引过程中须经过“概念—标识符号”的转换，使得分类标引的直观性较差，具有间接性。

(3) 由于主题法和分类法的体系结构不同，主题标引具有专指性、灵活性，分类标引具有系统性、稳定性。

主题法是直接以信息资源所研究的事物或对象为依据选择主题词，并可采用组配的方法描述主题，无论信息资源的主题如何专深，一般都可以直接选用主题词或通过主题词的组配来加以表达，所以主题标引具有较好的专指性。字顺系统是主题法体系结构的主体，即主题法系统是按照主题标识的字顺进行排列和组织的，它允许自己不断发展和增补，能对不断出现的新事物、新学科等的主题随时进行增补，灵活性较好。分类法是按学科性质划分的等级层累结构的逻辑分类系统，这种系统遵循从总到分、从一般到特殊、从低级到高级、从简单到复杂、从上位到下位，层层展开、上下隶属的逻辑序列，它能充分揭示事物之间严格的从属派生及平行的相关关系，使得分类标引具有较好的系统性和严密性。同时分类法又是一种难于随时改变类目体系的相对稳定的检索语言，尤其是普遍使用的学科体系分类法，因此一部分分类法编成后总是要求它保持相当时期的稳定性，这也使得分类标引具有相对的稳定性。

可见，主题标引和分类标引既存在共同之处，又有明显的差异。主题标引和分类标引的不同，主要是由两种不同组织方式的差异决定的，而且在不同系统中，主题标引和分类标引因采用的主题词表或分类表不同，其差异也是在变化的。

5.2.5 关键词标引

所谓关键词，是指那些出现在信息资源的标题摘要、正文中，对描述信息资源的主题内容具有实质意义的语词。如一篇题名为“网络媒体与传统媒体的融合”的文献，其中“网络媒体”、“传统媒体”、“融合”3个词可以有效地表达信息资源的主题，并具有检索意义，可作为关键词，而“与”、“的”两个词虽然能够帮助表达信息主题，但不起关键作用，也没有实际检索意义，因而不能作为关键词。关键词法是将信息资源原来所用的、能描述信息资源主题概念的那些具有实质意义的词抽出，不加规范或只进行极少量的规范化处理，按字顺排列，以提供主题检索途径的方法。关键词标引就是选择关键词对信息资源进行标引的方法。关键词标引一般通过计算机自动进行，即计算机自动抽取文献题名、文摘或正文中有检索意义的语词，通过轮排生成各种类型的关键词索引，如题内关键词索引、题外关键词索引、双重关键词索引、单纯关键词索引、词对式关键词索引等。

目前，在网络信息资源的组织和检索中大量使用关键词语言，如各种搜索引擎和网络数据库除了提供分类检索外，几乎都提供关键词检索途径，包括简单关键词检索和高级关键词检索（如布尔逻辑检索、加权检索、截词检索、字段检索、模糊检索等）。而且，由于网络

信息资源大量地以超文本文件、多媒体文件等非结构化文件形式存在,信息检索的智能化要求日益迫切,关键词语言在智能多媒体系统中也有了广泛应用。多媒体信息检索要充分利用文本、关键字和其他客观属性,综合其他学科领域的成果,并结合现有的关键词检索功能,集成到基于内容的检索系统中,利用特征之间的互补能力提高检索效率。

1. 关键词语言的类型

(1) 题内关键词索引。

题内关键词索引(KWIC, Keyword in Content Index),又称上下文关键词索引,是最早出现的机编索引,1960年首次用于美国化学文摘社出版的《化学题录》(Chemical Titles)。KWIC的编制过程是:选择文献标题中具有检索意义的词作为关键词(用禁用词表排除非关键词),并轮流排在索引款目中部的排检点位置,并保留关键词的上下文。如果文献标题过长,则以轮排的形式移至款目前部或后部,款目后跟随该信息资源的地址。如一篇题为《论网络时代图书馆的资源观》(文献号为8032)的文献,可通过计算机自动生成以下题内关键词索引款目:

检索入口

论	网络时代图书馆的资源观	8032
论网络	时代图书馆的资源观	8032
论网络时代	图书馆的资源观	8032
时代图书馆的	资源观/论网络	8032

(2) 题外关键词索引。

题外关键词索引(KWOC, Keyword Out of Content Index)是对KWIC索引的改进形式。由于KWIC将排检点设置在索引款目的中部,不符合一般用户的查找习惯,KWOC将索引标目的位置从中部移至左端或左上方,标目下完整列出文献篇名,原篇名中的关键词用特定的符号如“*”代替或予以保留。例如,上述8032号文献在KWOC索引中可生成以下款目:

网络	
论**时代图书馆的资源观	8032
时代	
论网络**图书馆的资源观	8032
图书馆	
论网络时代***的资源观	8032
资源观	
论网络时代图书馆的***	8032

(3) 双重关键词索引。

这是一种KWIC索引与KWOC索引的结合形式,即采用双重标目:在篇名之外设置第一个主标目,再在篇名的左端按副标目(第二关键词)排列。例如,上述8032号文献可编成下列双重关键词索引款目:

网络	
时代图书馆的资源观/论网络	8032
图书馆的资源观/论网络时代	8032
资源观/论网络时代图书馆的	8032
时代	
图书馆的资源观/论网络时代	8032

网络时代图书馆的资源观/ 论	8032
资源观/论网络时代图书馆的	8032
图书馆	
时代图书馆的资源观/论网络	8032
网络时代图书馆的资源观/论	8032
资源观/论网络时代图书馆的	8032
资源观	
时代图书馆的资源观/论网络	8032
图书馆的资源观/论网络时代	8032
网络时代图书馆的资源观/论	8032

2. 关键词索引编制的步骤

关键词索引编制主要由计算机自动进行, 其一般步骤如下所述。

(1) 由禁用词表控制抽词。

禁用词表亦称非关键词表, 是将那些没有实义或无检索意义的词如冠词、连词、介词、感叹词、代词及部分形容词、副词、动词和名词等非关键词编制起来形成的表。计算机自动抽词时, 首先通过禁用词表, 排除这些非关键词, 其他的词即可作为关键词。有些计算机抽词系统也可能编制关键词表, 即将相应学科专业领域中有检索意义的、可以作为关键词的词按字顺编排成表, 存入计算机内, 用以与题名、文摘、正文中的词进行比较, 将匹配的词作为关键词。这种关键词表一般编制比较简单, 不进行严格的词汇控制。

(2) 由计算机进行自动分词。

由于西文词汇之间有空格, 计算机对西文关键词的识别处理非常方便, 而汉语在书写时词与词之间不留空格, 计算机在切分一串连续的汉字字符时, 可能会有多种切分方案, 因此计算机自动编制中文关键词索引时涉及汉语切词问题, 目前国内采用的自动切词方法主要有词典匹配切分法、设立切词标志法、理解式切分法等。

(3) 通过轮排编制关键词索引。

关键词标引的结果一般是编制成关键词索引。通过轮排, 使每个关键词轮流排至检索入口, 编制成多条关键词索引款目, 提供多途径的主题字顺检索途径。

3. 关键词标引的评价

(1) 关键词标引的优点。

- ① 关键词标引时无须主题分析和查看词表, 简便易行, 降低了对标引人员的要求。
- ② 标引和索引编制易于实现自动化, 很多数据库关键词索引的编制已经实现了自动化, 大大缩短了检索系统信息组织的时差, 保证信息报道和传递的及时性, 适应了信息检索及时性的要求。
- ③ 关键词是信息资源中使用的自然语言, 表达主题比较直观、专指。
- ④ 所有的关键词都是平等的, 全部按字顺排列。从每一个信息资源中抽取的关键词, 每个关键词都是一个检索入口, 可提供多途径检索的功能。

(2) 关键词标引的缺点。

- ① 关键词标引直接采用自然语言作为关键词, 关键词基本上不进行规范化处理, 对自然语言中大量的等同关系词不加规范统一, 如对同一个词的单复数和变格等词形变化也不加统一, 而保持作者用词原状, 使相同主题的信息资源常常因作者用词不同而被分散, 导致漏检的可能性较大。

② 关键词标引不显示关键词之间的等级关系和相关关系,全部关键词在检索系统中彼此孤立,没有任何联系,增加了检全文献的难度。

③ 为了加速和简化检索工具的编制过程,关键词多限于从文献标题中抽取,但由于一些标题对信息资源内容的表达不够充分或不准确,会使关键词检索有一定的漏检和误检。

④ 在机编索引情况下,由于机械地抽词和轮排,其中有不少关键词款目是不起检索作用而徒增篇幅的。另外,汉语由于存在分词难的问题,应用计算机进行汉语关键词抽词标引时仍须解决词汇分词问题。

5.3 自动分类与自动标引

信息的自动分类与自动标引技术始于20世纪50年代,IBM公司的H.P.Luhn等发表了一系列文章,创立了信息自动处理研究领域。随着计算机及相关技术的飞速发展,该领域研究取得了令人瞩目的成果,并出现了许多实用化系统。

5.3.1 自动分类概述

所谓自动分类,就是利用计算机技术对信息(主要是记录型信息)按照一定的分类体系或标准进行自动分类标记,又可细分为自动聚类与自动分类两种。

聚类,就是根据信息内容的相关性来组织文献集合或信息集合,将整个集合聚集成若干个类,并使属于同一类的文档尽量相似,属于不同类的文档差别明显。

分类,即归类,将具有相近特征的检索对象相对地集中,而具有不同特征者尽可能归于不同的类别中。一般,分类是根据一个已经被标注(即分好类)的训练文档集合,找到文档特征和文档类别之间的关系模型,然后利用这种关系模式对新的文档进行类别判断。

对于自动分类,通常有两个基本要求:一个是分类结果应该与信息输入次序无关,即次序独立性原则;另一个是类别定义明确,即重叠度最小原则。

1. 自动聚类

聚类是一种重要的数据挖掘技术,可以用于从大量数据中寻找隐含的数据分布和模式。在文献信息处理领域,通过聚类可以将一批文献聚集成若干个类,提供一种组织信息资源的方法;可以作为一种文献信息分类的辅助技术,生成用于文本自动分类的分类体系表;可以发现与某文献内容相似的一批文档,以帮助用户获取相关知识。

聚类算法较多,常用的文本聚类方法主要有两类:等级聚类法与动态聚类法。

1) 等级聚类法

为表述方便,下面用“距离”指标来度量文档样本间或类间的相似程度,等级聚类的基本算法思想可以表述为以下操作步骤。

(1) 计算文档的距离系数矩阵。

首先,将待聚类文档集合中的所有文档(假设为 n 个)看做 n 个类,计算所有文档两两间的距离,共 $n(n-1)/2$ 个,形成文档的距离系数矩阵 $M_{n \times n}$ 。 $M_{n \times n}$ 是一个对称矩阵,其中,第 i 行 j 列的元素 d_{ij} 表示第 i 篇文档和第 j 篇文档之间的距离值,当 $i=j$ 时,很显然, d_{ij} 的值为0。

(2) 合并两个最相似的文档类。

从距离系数矩阵 $M_{n \times n}$ 的上(或下)三角元素中(对角线元素除外)找出距离的最小值,例如是元素 d_{ij} ,则文档 i_0 与 j_0 是最相似的,将它合并在一起,形成一个新类。这时,系数

矩阵中还有 $(n-1)$ 个类，记下参加合并的类的序号与距离值。

(3) 更新距离系数矩阵。

对于 $(n-1)$ 个类，重新计算两两之间的距离。由于矩阵中有 $(n-2)$ 个类没有变化，它们之间的距离也就不变，因此只需要对合并得到的新类与以前固有的 $(n-2)$ 个类两两间的距离值进行计算即可。类间距离的计算方法有多种，如最短距离法、最长距离法、中间距离法、重心法、离差平方和法等。

(4) 重复步骤(2)、(3)，直到所得文档类符合聚类要求。

根据上述步骤，每一次合并都使文档类的个数减少一个，经过 $(n-1)$ 次重复合并，原来的 n 个文档就聚合成一类。

实际应用中，通常用两种标准中断以上聚类过程：一是以聚类的个数作为标准，即当某一时刻文档类的个数为 k （ k 是事先规定的一个阈值）时就停止聚类计算；二是以相似度作为标准，即当某一时刻要合并的两个类的相似度低于某一给定阈值的时候，聚类算法就停止。

等级聚类法要对文档做全面的统计与计算，聚类比较准确，且聚类过程可视、可以追溯，但是当文档数量比较大时，算法的运行速度会比较慢。

2) 动态聚类法

较之于等级聚类法，动态聚类法试图避免全面的计算和比较，尝试在局部分析的基础上，对文档集合先做出某种较为粗略的划分，然后再按某种最优的准则进行修正，直到聚类结果比较合理为止，其流程如图 5.1 所示。

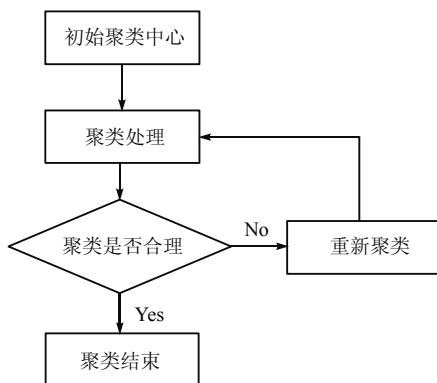


图 5.1 动态聚类法的流程

(1) 确定聚类个数 k ，从文档集合中选择 k 个文档作为凝聚点（即初始聚类中心），每个凝聚点文档自成一类。

(2) 按照距离最近原则，将剩余 $(n-k)$ 个文档逐个并入最近凝聚点所代表的类。每并入一篇文档，立即重新计算该类的重心，并用此重心替代原来的凝聚点。

(3) 以最后形成的每个凝聚点代表一类，将全部 n 篇文档重新聚类，逐个并入最近的凝聚点所属的类。与步骤(2)相同，每并入一个文档后，就重新计算重心，并以此重心代替原凝聚点。文档集合被重新聚类后，如果与前一次的聚类结果不同，就重复步骤(3)；否则，聚类处理即完成。

从以上流程可以看出，聚类个数 k 及相应的初始凝聚点的选择，不仅对聚类的最后结果有影响，也使得整个聚类过程具有很大的不确定性。

关于初始聚类中心的选取，有重心法、密度法、调用等级聚类法等方法。

重心法。首先计算出所有文档的重心，将其作为第一凝聚点；再选定一个正数，作为建

立新凝聚点的最小临界距离 d ；然后依次计算凝聚点与全部样本的距离，如果样本与已有凝聚点的距离大于 d ，则将它作为新凝聚点；再计算其他样本与新凝聚点的距离，大于 d 的样本作为新凝聚点；依此类推，直到所有样本处理完毕，就可以确定第一批凝聚点。

密度法。先选定两个正数 d_1 和 d_2 ，然后，以样本空间的每个样本点为中心，以 d_1 为半径做超维球，落在该球内的样本数（不包括中心样本在内）就称为该样本点的密度。选择密度最大的样本为第一凝聚点，再取密度次大的样本，计算它与第一凝聚点的距离，如果小于 d_2 ，就取消该点，否则选它作为第二凝聚点。按密度大小为序选取样本，用上述准则就可以得到两两距离都大于 d_2 的凝聚点集合。一般地，要求 $d_1 < d_2$ 。

调用等级聚类法确定初始凝聚中心。先利用等级聚类思想来寻找动态聚类所需要的 k 个初始凝聚中心，然后再进行动态聚类。这种方法可以避免初始聚类中心选择的随意性。

聚类参数 k 的选取。对于动态聚类而言，参数 k 既是聚类处理的起点，也是聚类结果的一部分。如果能使算法开始时指定的 k 值在聚类进程中进行适当的调整，以保持聚类结果的合理性，将不失为一个有效的方法。

k-Means 方法是一种典型的动态聚类法。基本的 k-Means 方法聚类操作较为简单：在聚类开始时，使用随机方式选择 k 篇文档作为初始聚类中心，按照前述动态聚类法的处理流程进行聚类处理，整个算法的结束条件是类的重心（或凝聚点），不再改变。

一般情况下，基本 k-Means 算法的聚类效果不如等级聚类好，而且由于初始的聚类中心来自随机选择的，产生的聚类结果非常不稳定。针对该算法，研究者提出了一些改进形式，如二分 k-Means 法。二分 k-Means 法的聚类思想：选择一个类进行分裂；用基本 k-Means 方法找出所选择类分裂出的两个子类（二分步骤）；重复二分步骤 m 次，选择其中能产生最大总体相似度类别的分裂方法；重复以上步骤，直到产生类的数目达到要求为止。

动态聚类法的优势是运算量小，能用于处理庞大的样本数据，也使实时处理有了一定的可能性。

2. 自动分类

所谓的自动分类，是在给定的分类体系下，根据信息的内容自动确定其所属类别的过程。在进行自动分类之前，已存在可以使用的分类表，这些分类表一般由相关领域专家事先制定，也可以通过聚类处理来获取。

文本分类是信息组织、检索与数据挖掘中的关键技术。一般地，根据一个已经被标注（即分好类）的训练文档集合，找到文档特征和文档类别之间的关系模型，然后利用这种学习得到的关系模式对新的文档进行类别判断。文本分类技术涉及很多预处理技术，包括中文分词、文本特征的抽取、文本的特征表示等技术，这些过程的质量与文本分类的最终分类质量有着密切的关系。

一般来讲，文本分类主要有以下问题需要解决。

（1）建立训练文档集合。训练文档集合应具有代表性，能反映分类系统所要处理的文档的情况。一般采用公认的、经人工分类的语料库。

（2）建立文档表示模型。建立文档表示模型是将文本数据转换为结构化数据的过程。向量空间模型是最常用的文本表示方法，在向量空间模型中，训练样本被表示成由特征项构成的向量空间。

（3）文档特征抽取。文本分类的一个重要难题就是文本的维度过高，这就需要采用特征抽取的方法，在不影响特征准确度的情况下，减少表示文本的特征，去掉一些信息含量少的词，以最终建立特征子集。

（4）选择或设计分类模型。选择分类模型实际上就是要使用某种方法，建立从文档特征

(或属性)到文档类别的映射关系,是文本分类的核心问题。

(5) 性能评测。性能评测是自动分类处理流程中的一个重要环节。完善科学的评测对改进和完善分类系统具有重要的指导意义。

现有的分类方法主要来自两个方面:统计和机器学习。常用的文档分类方法有 kNN、Naïve Bayes (NB)、向量空间模型 (VSM, Vector Space Model,) 法、支持向量机 (SVM, Support Vector Machine) 法等,下面介绍前两种方法。

1) kNN 分类法

kNN (k-Nearest Neighbours) 分类法由 Cover 和 Hart 于 1968 年提出,是一个理论上比较成熟的方法,该方法的思路非常简单直观:给定一个经过分类的训练文档集合,在对新文档(即测试文档或待分类文档)进行分类时,首先从训练文档集合中找出与测试文档最相关的 k 篇文档,然后按照这 k 篇文档所属的类别信息来对该测试文档进行分类处理。

kNN 分类法属于传统统计模式识别算法,其具体的分类过程如下所述。

(1) 对训练文档集合中的每一篇训练文档形成文档的向量表示,其分类情况则用一个分量值为 0 或 1 的类别向量表示。若类别向量的第 j 个分量为 1,表示此文档属于类 j ,若为 0 则表示不属于类 j 。

(2) 对于某一给定的测试文档 d ,通过计算文档之间的相似度,在训练集合中找到与其最相似的 k 篇训练文档,即 k 个最近邻居(用集合 kNN 表示)。显然, k 个最近邻居可能属于 m ($m \leq k$) 个不同的类别。

(3) 以每一个最相似文档 d_j ($d_j \in \text{kNN}$) 与测试文档 d 的相似度 $\text{sim}(d, d_j)$ 为权值,计算文档 d_j 属于类别 c_i 的决策规则值:

$$\text{score}(d, c_i) = \sum_{j=1}^k \text{sim}(d, d_j) \times y(d_j, c_i) - b_i d_j \in \text{kNN}; \quad i = 1, 2, \dots, m$$

在上式中, $y(d_j, c_i)$ 是文档 d_j 对类别 c_i 的分类值,其取值范围为 $\{0, 1\}$, $y(d_j, c_i) = 1$ 表示属于该类,而 $y(d_j, c_i) = 0$ 则表示不属于该类; b_i 为阈值,对于某一特定的类来说,其值是一个有待优化选择的值,可以通过一个验证文档集进行调整。

(4) 根据 $\text{score}(d, c_i)$ ($i = 1, 2, \dots, m$) 的值,最终决定测试文档 d 的类别归属。

这里,可以通过两种方法来判定文档 d 的类别。一种方法是:如果测试文档 d 只属于某一个类,则可以取 $\text{score}(d, c_i)$ ($i = 1, 2, \dots, m$) 中的最大值所对应的类别作为 d 的类别;另一种方法是:将所有的 $\text{score}(d, c_i)$ ($i = 1, 2, \dots, m$) 值进行排序,并指定一个阈值,测试文档 d 属于超过该指定阈值的所有类。

kNN 分类法不需要预先学习,其优点是分类精度较高,不存在漏识问题,缺点是分类速度与训练文档个数有关。因为,对于每一个测试文档,都必须求解它与训练文档库中所有文档的相似度,其时间复杂度为 $O(n_1 * n_2)$,这里 n_1 和 n_2 分别为训练文档总数和测试文档总数。

另外, $\text{sim}(d, d_j)$ 值的计算,如何兼顾文档特征之间的相互关联与共现,需要对算法进行改进。

最后, k 值的选择对分类结果影响很大。选取的 k 值较小时,选择的相关文档(或最近邻居)不足以表现待分类文档的所有内容特征;而选取的 k 值较大时,算法又会为待分类文档选择出一些不是很相关的文档,从而超出待分类文档的内容特征。在实际的文本分类试验中,要通过反复的实验、测试、观察,最后才能确定 k 的一个较理想的取值范围。

2) Naïve Bayes 分类法

Naïve Bayes 分类法 (以下简称 NB 法) 将概率模型应用于自动分类, 其分类思想是: 使用贝叶斯公式, 通过先验概率和类别的条件概率来估计文档 d 对类别 c_j 的后验概率, 以此实现对文档 d 的类别归属判断。

假设训练文档集合分为 k 类, 类别集合记为:

$$C = \{c_1, c_2, \dots, c_i, \dots, c_k\}$$

对于任意文档类别 c_i , 其先验概率为 $P(c_i)$, $i = 1, 2, \dots, k$ 。又对于任意测试文档 d , 其属于类别 c_i 的类条件概率为 $P(d/c_i)$ 。于是, 根据贝叶斯公式, 可得到文档 d 对类别 c_i 的后验概率为:

$$P(c_i/d) = P(d/c_i) \times P(c_i) / P(d)$$

对文档 d 进行分类, 就是按照上式计算所有文档类别在给定文档 d 情况下的概率, 概率值最大的那个类就是 d 应归属的类, 即当

$$P(c_i/d) = \max_{i=1}^k \{P(c_i/d)\}$$

时, 文档 d 属于类别 c_i 。对于给定的分类集合和测试文档, NB 法分类的关键就是计算 $P(d/c_i)$ 和 $P(c_i)$ 。换言之, 计算 $P(d/c_i)$ 和 $P(c_i)$ 的过程就是建立分类模型或分类器的过程。注意, 由于对于所有的分类过程, $P(d)$ 具有相同的值, 因此, 实际分类时可以忽略分母 $P(d)$ 的值。

首先是先验概率 $P(c_i)$ 的计算。通常, $P(c_i)$ 的计算方法较简单, 可以使用下式估计得到:

$$P(c_i) = c_i \text{ 类中的训练文档数} / \text{训练文档集合的全部文档数}$$

对于 $P(d/c_i)$ 的值, 计算比较复杂些。假设文档特征集合中的所有特征是互相独立的, 则有:

$$P(d/c_i) = \prod P(f/c_i) (f \in d)$$

f 为测试文档 d 中的特征。

令 n_{if} 为特征 f 在类 c_i 中出现的次数, n_i 为类 c_i 中的全部特征出现次数的和, $P(f/c_i)$ 有两种计算方法:

$$P(f/c_i) = n_{if} / n_i$$

和

$$P(f/c_i) = (n_{if} + 1) / (n_i + |V|)$$

式中, $|V|$ 为训练文档集合中全部不同特征的数目。

与其他文本分类技术相比, NB 法能很快地趋于稳定, 通常只需要扫描一遍文档即可完成分类处理, 速度较快, 因而比较适用于在联机环境下创建分类器。但是, NB 法假定在给定分类变量的情况下所有的特征项都是相互独立的, 这在文本分类的处理环境中显然是不现实的。

为此, 有不少针对 NB 法的改进, 使之能处理特征项之间存在有限相互关联 (或作用) 的情况, 即, 允许在 Bayesian 分类网络中, 一个特征节点除了类变量之外, 还有别的父节点。如此, 分类处理的复杂度将显著提高, 如 TAN、KDB 分类算法。

除了上述基于统计和机器学习的自动分类技术, 目前得到重视的还有基于知识库、知识

工程的规则分类技术。如侯汉清教授主持的“基于《中图法》的专家知识库系统构建研究”，采用以分类号（类目词）控制主题词，用主题词控制关键词，以分类体系为基础构建“分类—主题—关键词”一体化概念语义网络知识库，这种方法能在一定程度上降低分类难度，应用有待完善。其他一些规则分类方法，如基于粗糙集理论的分类规则生成算法、基于加权模糊推理网络的分类方法、本体论方法等，为建立高效准确的分类系统提供了新的思路。

5.3.2 自动标引概述

1. 自动标引的概念

自动标引（Automatic Indexing）又称计算机辅助标引（Computer Aided Indexing），是根据文献内容，依靠计算机系统全部或部分地自动给出标引符号的过程。换句话说，就是利用计算机系统模仿人的标引活动并自动生成情报检索所需的索引符号的过程。

相对手工标引，自动标引存在很大优势：速度快、一致性好、稳定性好、成本低，但准确率待提高。

2. 自动标引的原理

目前自动标引采用的理论主要来自三个方面：统计分析、语言分析、人工智能。

统计标引法是自动标引各种方法中历史最长的一种，是基于词汇分布特征的标引方法，也是目前较为成熟的一种方法。其理论基础是齐夫（Zipf）定律。

语言学家齐夫通过对文章中所用词汇的词频进行统计，总结得到：将一篇文章中所有词按词频的从高到低顺序排列，依次给出等级值 1、2、3……，则每个词的词频 f 与等级值 r 的乘积接近常数，这个统计规律被称为齐夫定律。

根据齐夫定律，可以在平面坐标系中得到一条 $f-r$ 的二次曲线（双曲线的一支），切分这条双曲线，可以把所有的词分为高频词、中频词和低频词。Zipf 定律是所有基于词汇分布特征的标引方法的基础。

高频词：传递信息能力弱，多为虚词，反映到文献标引上，为专指度小的泛指词，标引能力弱。

中频词：传递信息能力强，多为常用的术语，反映到文献标引上，则为标引时选词的最佳对象，专指度适中。

低频词：传递信息能力极强，产生的原因较复杂，可能是冷僻词，也可能是新引进的概念。反映到文献标引方面，这类词专指度太大，用自由词标引时可选取，若从词表中选主题词标引，则词表中无能力包括这类词，否则词表太大。

通常选取中频词和低频词作为文献标引的候选词，使用高频词建立停用词表。用停用词表方法，可在文献标引过程中极简便地排除高频词（泛指词），是表中的词则不作标引，故称停用词表。

语言法中，通过对构成文献的自然语言的分析，利用一定算法产生标引词，是基于语言规则与内容的标引。这是从语言学角度对自动标引方法的探索。主要包括两种方法：句法分析、语义分析。

人工智能法是指利用计算机从事标引工作中的脑力劳动，即让计算机模拟标引员完成标引文献的工作。基于人工智能的分词方法可分为专家系统分词法和神经网络分词法，是目前中文分词技术的一个重要研究和发展方向。

3. 自动标引的过程

自动标引的一般流程如下所述。

(1) 形成机读形式的文献。将待标引文献按标引系统要求的格式输入系统,针对印刷型文献,可以通过手工录入或光学字符识别(OCR)输入;电子文档(xml、doc、txt等格式)可以直接导入,才可能进行自动标引。

(2) 语句分析。借助一定的技术手段对机读文献中的语句进行分析,区分词与非词、实词与虚词。

(3) 语词加权。设计或确定实词的加权方案,并根据该方案计算每个词的权重。

(4) 确定标引词的阈值。根据待标引文献的标引深度要求,兼顾各种文献的特性,确定候选标引词的词权阈值。

(5) 选出标引词。根据确定的阈值选出词权不小于阈值的候选词作为标引词。

(6) 将标引词转换为受控词。将选出的标引词转换为词表中的受控词。使用关键词-受控词对照表,根据其中关键词与规范化的主题词、副主题词、特征词之间的对照关系,进行对应转换;或者利用词汇相似度,大多数意义相同或相近的词的字全部或部分相同,关键词与主题词之间存在一定程度的相似性,可通过某些算法计算出来,根据相似性确定相应的主题词。

(7) 生成索引文档,输出索引。根据前面确定的全部标引词连同它们的地址信息,按照某种要求自动组织排序,生成检索用的倒排档或词典文档,完成自动标引。

自动标引质量的改进离不开用户在检索过程中的反馈,注意跟踪用户的相关性判断,进行词加权计算或改进。

4. 自动标引的类型

(1) 按人工介入与否,自动标引分为全自动标引与半自动标引。

全自动标引:包括内容分析、词的识别和主题表达等方面的标引作业各环节全部由计算机完成,完全或基本上不需要人工干预。

半自动标引:全自动标引中的某一个或几个步骤由人工完成,也称计算机辅助生成索引系统。

一般情况下,若不加说明,自动标引仅指全自动标引。

(2) 按标引词来源,自动标引分为自动抽词标引与自动赋词标引。

自动抽词标引:指利用计算机直接从原文(即文献题名、文摘或正文)中抽取关键词作为标引词,并自动生成关键词索引或倒排档的过程。又分为主关键词索引和全关键词索引,前者是在后者的基础上再选出少量主要关键词作为标引词。

自动赋词标引:指使用预先编制的词表中的词来代替文本中的词汇进行标引的过程,也就是将反映文本主题内容的关键词(欲用于标引的关键词)转换为词表中的主题词(或叙词等),并用其标引的方法。这是计算机模拟人的赋词标引方法。有基于概率的赋词标引与基于概念的赋词标引两大类。赋词标引都是在抽词标引的基础上实现的。

5.3.3 西文信息自动标引技术

西文是指拉丁字母的拉丁语系,包括英文、法文、德文、西班牙文、葡萄牙文等,不包括俄文、日文等。目前在国际上,西文文献,无论是纸质的还是网络型的,在数量和质量上都占有极大的比例。

1. 基本标引技术

自动标引技术在西文信息的处理方面已基本成熟和实用化。在标引过程中,需要解决的主要问题是:抽取关键词、分析确定标引词。

1) 抽取关键词

利用计算机抽取西文关键词, 需要完成的是从文本中剔除虚词(又称非用词或停用词), 获取关键词。这就需要在抽词之前, 首先建立一个以介词、冠词、连词等无实质意义的单词(如 at, for, the, and, of 等)组成的非用词表(Stop-List)。然后利用非用词表, 从被标引的文本中筛去非用词, 抽取关键词。

取词的一般过程为:

(1) 从待标引文本中抽取一个单词, 由于西文中每两个单词之间都具有空格间隔, 因此, 可以遇空取词, 这一点较之于中文, 方便得多;

(2) 确定关键词, 利用取出的词去搜索非用词表, 是非用词则舍去, 是关键词则记下;

(3) 分析关键词, 重复关键词, 累计词频, 如果标引对象为全文, 还可以根据位置赋予权重。

2) 分析确定标引词

标引词的确定, 主要依据词频统计及其他一些途径。这里, 词频包括: 绝对词频与相对词频。绝对词频指的是词在一篇文献中出现的频率; 相对词频则是指将词在一篇文献中的出现频率与在整个文献库中出现的频率进行比较。

在进行词频统计时, 可以根据从不同位置取出的词给予不同的权值。例如, 可以对标题词给予最高权值, 其余的可以从大到小依次对文摘词、首尾段词、首尾句词等赋予权重。所取出的关键词是否作为标引词, 可根据计算每个被抽取词的权值之和, 按从高到低的顺序确定。若取词对象是标题, 只需判断所取出的词是否重复。若从文摘或全文中取词, 则需根据词频统计的结果去除低频词, 将高频词作为标引后备词, 根据系统规定的标引词的数量, 最后确定标引词。

2. 标引词加权方法

1) 词频统计标引法

词频统计标引法也称为绝对频率加权法, 这是一种根据词的出现频次来确定词的重要程度的标引词加权方法。其主要步骤如下。

(1) 给定 n 篇文献组成的一个集合, 计算每篇文献中每个不同的词的出现频率 f_{ik} (词 k 在第 i 篇文献中出现的频率)。

(2) 计算每个不同的词在整个文献集中出现的频率, 得到各词的集合频率: $f_k = \sum f_{ik}$ 。

(3) 按照 f_k 的大小将词降序排列, 用试错法确定高频词和低频词的阈值。确定一个上截止阈值, 去掉 f_k 大于上截止阈值的词, 因为这些词往往是一些仅起语法作用的功能词或内容很泛指的词; 确定一个下截止阈值, 去掉 f_k 小于下截止阈值的词, 因为这些往往是文献集中的罕见词, 区分不同文献的能力较差。

(4) 去掉高频词和低频词后, 将余下的中频词选为标引词。

这种方法的特点是: 简单、容易实现, 有一定的实用性。但是简单地排除全部高频词和低频词, 可能会降低查全率与查准率, 且确定上下截止阈值也存在一定的困难; 更重要的是, 词频并不能全面涵盖词在文本中的功能。这也正是该方法的缺点所在。

2) 逆文献频率加权标引法

逆文献频率加权标引法基于如下假设: 某词的重要性与它在特定文献中出现的频率成正比, 而与该词在整个文献集中出现的频率成反比。

设 F_{ik} 为词 k 在文献 D 中的出现频率, DF_k 为包含词 k 的文献数, 称为词 k 的文献频率,

即

$$DF_k = \sum_{i=1}^n df_{ik} \quad df_{ik} = \begin{cases} 1 & F_{ik} \geq 1 \\ 0 & F_{ik} = 0 \end{cases}$$

词的出现频率只对文献集中某确定的文献才有意义,而词的文献频率则是对整个文献集合而言的。在一个文献集中,非特征词的文献频率一般较高,如“的”、“地”等反映句子语法结构的词,几乎在所有文献中都出现;而特征词的文献频率一般较低,如“超导”一词通常只在一些主题内容与超导有关的文献中才出现。

在一篇特定的文献中,特征词的出现频率越高,说明它与该文献的主题相关程度越高。所以在标引中,人们总希望所选择的标引词在某个特定文献中的出现频率较高,而在整个文献集合中的出现频率较低。一个词如果文献频率较低,说明它不是特征词,若这个词在某篇文献中的出现频率较高,则这个词可以较好地反映该文献的主题内容。因此在文献频率一定时,词的出现频率越高,越能较好地揭示文献的主题内容,即高频特征词是较好的标引词。在设计标引词权重时,其大小应与标引词的出现频率一致,与标引词的文献频率成反比。根据这一思想,标引词权重设计如下:

$$W_{ik} = F_{ik} / DF_k$$

此式说明,对于一定的词出现频率 F_{ik} ,标引词的权值随文献频率 DF_k 的增大而减小,随 DF_k 的减小而增大,即标引词的权重与标引词的文献频率具有互逆关系。因此这种标引称为逆文献频率加权标引。

要说明的是, DF_k 也可用于代表词 k 在整个文献集中的词频的总和,这样该方法等同于文外频率加权法。

3) 词区分值加权标引法

词区分值描述了词的区分能力,即词对文献的“分离”能力。如果一个词能较好地反映出文献集中各文献的差异,则这个词区分文献的能力就较强。因此,可以从词区分文献的能力出发来设计标引词权重。

设有 n 篇文献构成的集合 D ,第 i 篇文献表示为 $D_i = (d_{i1}, d_{i2}, \dots, d_{it})$,式中, d_{ij} 为文献 D_i 的第 j 个标引词的权值,则该文献集合的矩心 C (在自动分类中,矩心也称做类目中心)为:

$$C = (Cd_1, Cd_2, \dots, Cd_t)$$

式中

$$Cd_k = \frac{1}{m} \sum_{i=1}^m d_{ik} \quad k = 1, 2, \dots, t$$

将空间密度 Q 定义为所有文献与矩心相关程度的总和,即

$$Q = \sum_{i=1}^n S(C, D_i)$$

式中, $S(C, D_i)$ 为文献 D_i 与矩心 C 的相关程度。

又设, Q_k 为去掉第 k 个标引词(也就是 t 维向量变成 $(t-1)$ 维向量)后的文献空间密度,则词 k 的区分值定义为:

$$DV_k = Q_k - Q$$

如果一个词的区分值大于零，则用其做标引词会使文献间的相似度减少，使文献空间密度降低，从而使标引效率提高，因而设计词权时应取较大的权值；如果一个词的区分值小于零，则用其做标引词会使文献间的相似度增加，使文献空间密度增大，从而使标引效率将低，因而设计词权时应取较小的权值。也就是说，标引词权重应与标引词的区分值成正比。根据这一思想得加权函数如下：

$$W_{ik} = F_{ik} \cdot DV_k$$

4) 词相关性加权标引法

概率检索理论认为：最好的标引词是那些趋向于出现在与某一提问相关的文献中的词。

在文献集合 D 上给定提问 $Q = (t_1, t_2, \dots, t_m)$ ，设初始文献标引采用未加权的二值标引系统， Q 中词向量元素 t_k 所对的标引词 k 出现与否所得到的检索结果如表 5.1 所示。

表 5.1 词 k 与检索结果的关系

词 k 状态	相 关 文 献	不相关文献	合 计
词 k 在文献中出现	r_k	$n_k - r_k$	n_k
词 k 在文献中不出现	$R - r_k$	$N - n_k - R + r_k$	$N - n_k$
合计	R	$N - R$	N

表 5.1 中， n_k 为含有词 k 的文献总数， r_k 为含有词 k 的相关文献数， N 为文献集合中的文献总数， R 为与提问 Q 相关的文献总数。

假设词的分布在所有相关和不相关文献中均是独立的，相关性仅取决于检索词 k 在文献中出现的概率（频次）。根据表 5.1 和假设，Salton 等人利用概率推导法得到了加权函数：

$$W_{ik} = F_{ik} \cdot \frac{r_k(R - r_k)}{(n_k - r_k) / [N - n_k - (R - r_k)]}$$

这种方法的缺点是 R 和 r_k 的值难于准确得到，也就难以计算相关值。

关于标引词加权的方法还有很多，如统计学习标引法、概率标引方法等。统计学习标引法首先通过学习过程建立候选标引词与对其标引产生正反不同作用的促进词和削弱词集合之间的关系，然后由标引过程根据候选标引词在此关系中的权值及其词频来确定其是否作为标引词，这种方法由学习和标引两个过程组成。概率标引法所依据的概率主要有相关概率、决策概率和出现概率。基于相关概率的标引法是根据包含相同标引词的提问与文献的相关概率来标引划分文献的；基于决策概率的标引法主要依据某标引词赋予某文献这一决策事件正确的概率来标引文献；基于出现概率的标引法根据词在文献中的出现频次所服从的概率分布的特征来选择标引词。

5.3.4 汉语信息自动标引技术

1. 汉字的性质与特点

汉语是一种象形会意语言，字与字之间、词与词之间的组合灵活多样，且字与字之间、词与词之间没有明确的分隔标记，加上汉语词汇存在着一词多义、多词一义等现象，这些特点给汉语语词切分和词频统计带来许多困难。其中，最主要的难题是语词切分，若解决了这个问题，其他的标引过程与西文标引技术类似。本节讲的汉语文献自动标引方法，其实大部分讲的是汉语分词算法。

中文语词切分的困难主要源于以下几个方面。

(1) 汉字组词方式复杂, 如不联系上下文, 很难正确分词; 例如, “发展中国家兔的饲养”一句, 现有的汉语就可能导致两组语词分隔结果: 发展中国家/兔/的/饲养; 发展/中国/家兔/的/饲养。

(2) 交集型的标引词汇难以处理; 例如, “并行程序设计语言”为一文献名称, 其标引词应为: 并行程序设计/程序设计语言, 而不应从任意一处简单分开。

(3) 汉语虚词众多, 而且绝大多数汉字与不同的汉字组词时, 可能为关键词, 也可能为非用词; 如, “非”与“是”、“常”、“洲”可分别组成不同意义的词“是非”、“非常”(非用词)和“非洲”(关键词)。

(4) 新词频繁出现, 如人名、地名、新产生的词、外来词等, 这些都给汉语分词增添了难度。

虽然汉语词的切分存在很多很大的难度, 但在中文信息处理需求快速增长的驱动下, 通过众多研究者多年的努力, 已取得了很大进展, 下面将介绍几种汉语词的切分技术。

2. 汉语信息的切分标引

切分标记法, 就是将能够断开句子或表示汉字之间关系的汉字集合组成切分标记机内字典, 这个字典称为切分标记字典。

汉语信息原文由若干句子组成, 而句子之间由标点符号分隔, 每个句子由若干词、词组或短语组成。计算机若能获得句子中短语或词组间的分隔标记, 就能进行语词切分。切分标记分词法的出现正是基于这样的思路。

这种方法的优点是: 无须构造词典, 只须构建一个规模很小的标记字典。切分标记字典包括: 词首字、词尾字、独立字或几种情况的组合字, 也有以“非用字”、“条件用字”等组成切分字典的。利用切分字典将待标引文本中的句子分割成汉语词、词组, 再按一定的分解模式将它们分割成单词或专用词。实践表明, 字典完全可以替代词典完成自动标引。当然, 构造一个能用于切分的字典也需要一定的汉语语言知识和专业知识, 汉语语言的丰富复杂和专业词汇的纷繁各异, 给字典的构造带来了一定的难度。

非用词后缀表法是切分标记法的典型代表。该法将汉字分为非用词、条件用词、表内用词、表外用词。

非用字: 单字本身及与其他后缀字所组的词均不能作为标引词的字, 如“的”、“其”、“如”等。

条件用字: 单字与一些后缀字组词时可作为标引词, 与另一些字组词时则不能, 如“情”、“轻”、“使”字。

用字: 单个汉字也能作为关键词的字, 如“电”、“光”、“国”等。

表内用字: 单字本身是用字, 与其他词组又可以构成新的标引词的字。

表外用字: 非表内字均视为表外用字。

前三类字置于“非用字后缀表”中, 并给出区分标记, 如表 5.2 所示。

表 5.2 非用字后缀表示例

键 字	标 识	后 缀 字	键 字	标 识	后 缀 字
啊	0		色	1	列盲 ...
半	1	导岛 ...	甚	0	
测	1	量 ...	声	2	控音 ...
的	0		算	1	法机 ...
电	2	波流信压阻子 ...	虽	0	

续表					
键 字	标 识	后 缀 字	键 字	标 识	后 缀 字
管	1	理 ...	所	0	
光	2	电合纤 ...	通	1	道通讯 ...
分	1	析子 ...	信	2	道息源...
计	1	算	颜	1	色 ...
雷	2	达电鸣 ...	以	1	色
情	1	报 ...	在	0	
然	0		之	0	
如	0		至	0	

注：非用字、条件用字、表内用字分别以 0、1、2 标识

利用非用字或非条件用字进行分词时，根据“有联系则取，无联系则断”的原则，字间有无联系，主要根据条件用字和非用字来进行判断。

非用词后缀表法有一定实用性，但其算法复杂，字典构造困难，其使用的普遍性远不如词典法和单汉字标引法。

3. 词典切分标引法

词典切分标引法是借助词典抽取文献中词汇，并进行标引的过程。这种方法是目前汉语自动标引算法中占比重较大的一种。常用的有主题词表法与部件词典法。

- 1) 主题词表法
- 主题词表法是以主题词表为主，辅以禁（停）用词表和逻辑判断规则的标引方法。其基本标引过程如下。
- (1) 利用禁用词表排除输入的文献题目中的禁用词，并将剩下的短语记录在短语文件中。
 - (2) 利用机读主题词表对短语文件中的短语逐一比较（正向最长匹配、正向最短匹配、逆向最长匹配法、逆向最短匹配法），抽出匹配相同词并将其所在位置、范畴号和词族等信息记录在抽词文件中；其中常见的选词算法有：正向最长匹配、正向最短匹配、逆向最长匹配、逆向最短匹配等。
 - (3) 利用汉语的局部语法特征和一些主题判断规则对上述两种文件的信息进行加工，确定用于标引的主题词。
- 2) 部件词典法
- 部件词典法是以建立一个“二字部件词典”和一个“一字部件词典”为基础的标引方法。所谓“部件词典”，就是由许多“部件词”及其“词性”组成的表。部件词可以是整词或词的词头、词中、词尾，也可以同时兼有不同部位。因此，根据不同的组合，部件词可以产生 15 种状态（词性）。为了处理上的方便，每一种词性或组合词性都用一个数字编码唯一确定，参见表 5.3。

表 5.3 部件词典词性表

编码 词性	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
独立	☆		☆		☆		☆		☆		☆		☆		☆
词头		☆	☆			☆	☆			☆	☆			☆	☆
词中				☆	☆	☆	☆					☆	☆	☆	☆
词尾								☆	☆	☆	☆	☆	☆	☆	☆

每个部件词,必定对应于表 5.3 中 15 个词性中的某一个。如:某部件词既可以做词头,又能出现在词中,则它的词性为 6;某部件词只能出现在词尾,则它的词性为 8。

汉语的词汇相当丰富,如果要建立一个涉及各个领域的词典,则词典规模将十分庞大,维护管理相当困难,而构造一个二字部件词典和一个一字部件词典则相对容易得多。因此,采用部件词词典来代替关键词词典,这样不仅可减轻组织管理负担,而且可提高处理速度。

在部件词典中,每个部件词都有一个对应的词性状态值(大于等于 1,小于等于 15)。部件词典法标引的基本过程如下。

(1) 采取顺向或逆向扫描方式对文献进行扫描处理。

(2) 找出当前的二字,查“二字部件词典”,看其是否为部件词。

(3) 若是部件词,记下,继续查以下的二字;若不是,则退位找到当前二字的第一个字。查“一字部件词典”,看该字是否为部件词,若是则记下,若不是则从第二个字开始,以二字查找。

(4) 按上述方法依次对文献语句进行扫描,查出全部部件词,然后按 15 种词性进行组配,生成整词。

(5) 可用如上组成的整词进行标引,也可对整词进一步规范化处理,得出最后的标引词。

词典法是一个传统的标引方法,因其算法较为简单和清晰,因此在汉语自动标引中使用得相当普遍。但是词典的构造较困难,词典的维护、更新代价大,词典法学习新词能力差。

在实际应用中,出现了基于基本分词词典的统计分词方法,也称为概率分词,通过字与字相邻共现的频率或概率来反映成词的可信度,也就是对语料中相邻共现的各个字的组合频度进行统计,计算共现概率,超过一定阈值的,则认为此字组可能构成了一个词。这种方法既发挥匹配分词切分速度快、效率高的特点,又可以通过统计方法识别生词。

4. 其他汉语信息标引技术

除了前面提到的几种自动标引方法,还有一些其他方法,如单汉字标引法、语法分析标引法、神经网络分词法等。

1) 单汉字标引法

单汉字标引法以单汉字为处理单位,利用汉字索引文件实现自动标引和逻辑检索。由于这种方法把对“词”的处理改为对“字”的处理,因此就绕过了汉字分词的难题。由于每个汉字都由计算机做索引,无须人工标引,因而单汉字标引也被称为“无标引”。从计算机内部来说,对每个汉字均做标引词处理,因而也可理解为“全标引”。

单汉字标引和检索的基本过程是,对要处理文本逐一抽字,经过一些处理(如去掉无意义的虚字)后,建立索引文件。检索时将检索词拆分成单字与索引文件进行比较,并运用逻辑组配得出检索结果。

单汉字标引技术自 20 世纪 80 年代中期提出以来,算法得到了不少改进。

(1) 引入西文全文检索中的“位置运算”概念,在索引文件中,以单汉字为单位,记录其在各篇文献中的位置信息。在检索时,将检索词拆成单字,查找单汉字索引,然后根据单字的记录号与位置集合进行记录号的匹配,从而得到命中集合。随后又提出“串检索”的概念,在此基础上,形成了“首字定位,全词匹配”的检索算法。

(2) 在单汉字检索系统中增设后控词表,帮助检索者获得有关的同义词和相关词,以辅助与加强检索功能。对后控词表进行改进:以主题词表为骨架,附加以全面的自由词,并标示出包括所有主题词、自由词的用、代、属、分、参等关系,其具体实现是为每一个主题词、自由词赋予一个词号,在词表中每个词的入口项后分别标出其同义词、相关词及上、下位词的词号。可以全面矫正用户检索词,深层次提示隐含主题,增强了系统的扩检、缩检能力。

另外，还有不少优化技术，有采用加注标引方法，弥补后控词表的不足；检索结果、主题词文档自动生成结果文档，加速系统响应速度；等等。

单汉字标引和检索技术是完全的后组式标引和检索技术，绕过了自动分词，具有极大的灵活性；另外，单汉字系统丢失了以词为基础的检索系统中具备的许多重要信息，给机器运行和用户智力带来了额外负担。

2) 语法分析标引法

语法分析标引法通过对自然语言文法或句型文法的分析来抽取主题词加以标引。

由于汉语自然语言文法复杂，规则较多，目前还没有一个形式化系统能对汉语文法进行描述。目前的语法分析标引法并不是要分析全部汉语文法，而是专门从文献中挑选形如“本文讨论了……”这样的特征句型。因为这些句型正是表达文献主题内容的句型。因此可以用这些句型抽取主题词进行标引。所以，严格地说，目前这样的语法分析标引法只能称为“句型文法分析法”，而且只适用于科技文献。

识别出这类特征句子后，由自动抽词处理器对句子进行抽词。得到候选标引词。候选标引词经过禁用词表处理后，便进行加权处理。

加权的方法有：

(1) 对句型文法中每一句型赋予权值，当候选标引词从某一句子中抽出时，该句型的权值就是候选标引词权值的一部分；

(2) 根据已存在的标引词的统计特性来确定权值，即这一标引词在以前标引过程中平均的权值；

(3) 在切分处理中，如果一对候选标引词相匹配，则它们的权值有所增加。

将整个过程所得的权值相加，就是某一候选标引词的权值，然后根据设定的阈值判别其是否作为标引词。

3) 神经网络分词法

神经网络是在模拟人脑结构和行为的基础上，用大量简单的处理单元广泛连接组成的复杂网络。基于神经网络的分词算法将分词知识的隐式方法存入神经网络内部，通过自学习和训练修改内部权值达到正确的分词结果。分词关键在于知识库、权重链表的组织和网络推理机制的建立。

当前，基于神经网络的分词研究主要侧重在解决分词歧义上和未登录词识别两个难点问题。神经网络分词流程如图 5.2 所示。



图 5.2 神经网络分词流程

从分词样本中取出句子，对其进行编码压缩，变换成神经网络能够识别和存储的编码形式，然后提交到输入模块的入口，直到没有待切分的语句。

神经网络模型中的隐含神经元表示一组关联规则，输入的一组编码对应于关联规则的条件，而输出规则对应于关联规则的结果，也就是待切分语句的切分。具体而言，对于刚初始化的神经网络分词模型，可以先输入一定数量的样本进行训练。输入层每一个神经元均对应固定的字或词，每个样本都有其自身的切分规则。可以把这些规则理解为网络的权重，一旦训练完成，系统就能对由这些字词做出正确切分，使神经网络实现自适应和自学习，以获取

新的知识。汉语歧义和规则非常复杂,神经网络的学习过程是一个循序渐进的过程,加大训练次数,可以使切分词语的精度得到提高。神经网络是一个动态学习的过程,在已得到训练的神经网络中,如果以后输入的语句与原有的切分规则近似,则可以输出与样本近似的切分结果;如果以后输入的语句与原有的规则有较大差别,则神经网络把它看成新的切分规则进行学习。

神经网络的每一个输出节点代表一种切分方式,可以运用已有的知识切分汉语,但对自身的工作结果不能做出明白的解释,必须用一个输出模块来解读网络生成的结果,这个模块属于系统分词输出的后处理。

5. 几种标引方法的比较

一直以来,国内语言学界、人工智能领域和情报检索界的学者对汉语自动分词这一研究领域给予了极大的关注,提出了许多解决汉语自动分词的方法,这些方法各有优劣。下面对几种标引方法做简单的对比分析。

词典切分标引法与切分标记法属于先组式标引,检索无须对字串中的字间关系进行组配,检索速度快。基于词典的分词方法由于其算法成熟,易于实现,是目前普遍使用的切分方法;词典切分标记法中的主题词表法具有的扩检与缩检功能其他方法望尘莫及的。其局限在于:词(字)典的构造较为繁重,且需要及时维护,否则很容易滞后。

单汉字标引法避开了分词的障碍,实现容易,不存在词典构造问题,新概念词能及时处理,解决了汉语交集型字符串标引的问题。但是单汉字标引法对文本中隐含概念难以处理,容易造成漏检;运算复杂,与先组式检索系统相比,这种处理技术比较耗时,降低了系统的检索速度;而且无检索意义和分辨率低的词占有很大比例,易造成严重的空间浪费。

语法分析标引法在理论上比单纯以词典和字典为基础的标引法要深入完善得多,分词效果也优于前者。但是由于汉语的语法复杂,这种方法在实践上的发展还较缓慢,目前在自动标引中出现的相对简单易行的仅有句型文法分析法,且标引对象大多限于科技文献的标题和文摘,因为在这些文本中句型数量有限,变化不大,易于归纳和描述,要将分析范围扩大到全文和其他非科技领域,还有很多工作要做。

神经网络分词法是目前理论上理想的分词方法,但是目前基于神经网络的分词研究采用的样本集大多是小范围内的文本,知识表示规则不全面,算法本身存在训练时间长,收敛速度慢、知识表示复杂等,进一步提高该模型在分词领域中的实用性和分词效率,对于中文信息的自动化处理具有非常重要的意义。

在实际应用中,各种分词方法可以优势互补,结合使用,对于任何一个成熟的分词系统,综合不同的算法是必须的。

5.3.5 信息自动标引有待研究的问题

自动标引技术的提出,最初源于解决文献缺少关键词的问题,发展至今日,其应用早已超过这一范围。目前自动标引技术已经被广泛用于信息检索、自动问答、数据挖掘等领域,并且随着网络信息量的迅速增长及信息处理技术的不断提高,这个传统的研究课题将不断地被赋予新的含义和任务,其应用领域也将不断扩展。要使自动标引技术适应发展要求,对以下问题应该进行深入研究。

1. 主题词标引

主题词标引,较之于其他标引方法,具有标引规范、良好的扩检与缩检功能等优势,因而得到了普遍应用。但是传统的主题词标引,基本是先取得关键词的权重,排序后立即输出

若干权重相对较高的词，限于标引词的数量，最终的标引结果可能无法概括文章的主题。另外，主题词的不连贯性也常使检索者难以直接从主题词中准确地揣摩出文章主题。因而有研究者提出了主题概念标引：通过选取直接上位词作为主题概念、聚类产生上位词作为主题概念或将两个（或以上）主题词合成生成主题概念等方式确定概念词。能正确进行主题概念标引的关键问题是设计一部好的层次概念词典，这也就成为研究的主要内容。

要真正实现概念层次的自动标引，需要借助语义网技术。在主题词自动标引的过程中，结合主题词表的分类体系构建语义网。在自动标引过程中，对于所得主题词，在语义网的框架下进行语义逻辑推理，得到具有语义意义的标引词。基于本体的自动赋词能够在概念层面对文本进行标引，并能识别概念之间的关系。随着本体构建技术的发展，将其应用于语义检索，应该是有意义的探索。

目前，无论是科技领域还是商务领域的从业者，对信息的需求往往十分具体，有时会细化到要求得到某个统计数据、工艺参数或销售地点、数据。可以想象，这些信息可能只存在于文中的某个段落甚至某句话中，且只出现一次，极有可能被标引系统或标引人员所忽略。再者，传统的信息标引，其标引深度有限，会造成信息遗漏。为了解决标引深度与需求之间的矛盾，出现了关键词标引，但是关键词的不规范性会影响检索效果。如果能建立基于全文的主题词标引，将是提高检索效率的一种有效途径。尽管基于全文的主题词标引技术是汉语文献主题标引的一个难点，但是如果能够加大投入，加强研究，这一技术的实用化并非没有可能。

2. 加权标引

标引过程中的一个关键步骤是给可能用做标引词的词赋予权值，然后依据权值对这些词进行筛选，确定哪些词可以用做标引词，因而词加权方法对标引质量有至关重要的影响。

以统计方法为主的权重处理方法中，普遍考虑的是词频和词的位置这两个属性。利用词频进行加权的线性加权模型是文本检索权重计算常用的模型，该模型结构简单，使用方便，在文本检索、文本分类等领域广为应用。但是主题词在文中出现的次数与其在表现文献主题内容重要性之间简单的线性关系，难以令人信服。另外，尽管特征词在不同的位置对其相应的权重应该有所影响，但是很难将这种影响表示为明确的定量关系。

目前，文献主题概念的抽取与选择已经成为用计算机进行自动摘要、自动分类等涉及对文本信息理解进行智能处理工作的基础。目前，在进行主题概念选择时，主要考虑的因素有概念频统计、位置加权、概念长度、概念类型等。研究者提出了文献映射主题词选择方法、非线性加权体系、基于概念层次树的主题词轮排选择算法、基于关联矩阵的主题概念选择算法等。后两种方法考虑了主题词之间在概念上的相关性，能更有效地实现主题概念的标引。这些标引算法的优势在实验过程中得到了证实，是否能够投入使用，有待进一步的研究。

随着网络的普及，Web 文档的标引引起了关注，Web 页面本身带有标记，权重设计必须考虑这些标记。一些研究者结合位置加权与标引源加权方案，对 Web 文档进行标引试验，取得了良好的效果。适应网页动态性等特点，设计出合理的权重计算方案，对网络信息的检索与挖掘具有重要的意义。

3. 检索技术的革新

标引的目的在于检索，因此标引技术的发展与检索技术的发展是密不可分的，检索技术的革新必然会对信息标引提出新的要求。

信息检索技术经历了手工检索、脱机批处理检索、联机检索至网络化检索的过程。随着信息技术的快速发展和信息用户检索需求的增长，现代信息检索技术的发展将会呈现出智能化、专业化、可视化、跨媒体的趋势，最终在语义、语用、语境层次上实现智能化的信息检

索。在此过程中,需要深入研究的技术包括:智能化信息检索技术、可视化信息检索、专业化信息检索、跨媒体搜索等。

智能化信息检索技术是基于自然语言的检索形式,机器根据用户所提供的自然语言表述的检索要求进行分析,形成检索策略进行搜索。系统能够代替或辅助用户完成诸如选词、选库、构造检索式,甚至在数据库中进行自动推理查找等功能,可以使用户得到能够直接加以利用的信息。这意味着用户将从烦琐的规则中解脱出来。

可视化信息检索,是指用户在检索过程中各检索对象之间的关系以可视化的形式展现在用户面前,用户顺着可视化的检索画面一步一步地发现检索结果,检索结果也以可视化形式呈现。这种技术缩短了用户理解信息的时间,提供了感觉与思考之间的有效反馈机制。尽管目前成熟的、商业化的可视化信息检索系统还未问世,但随着网络技术的发展,可视化检索的优势将越来越突出。

专业化信息检索是指面向某一特定专业或学科领域,提供高质量的专业信息的检索。专业化信息检索需求的出现主要是因为网络信息资源越来越丰富,而综合性检索系统如搜索引擎查找专业信息越来越困难,效率比较低,往往不能检索到高质量的专业信息,发展专业化检索将是未来的一个研究热点。

跨媒体搜索,指的是通过多种媒体源之间语义关联分析和融合,允许用多种媒体信息表达用户检索需求,并最终输出多种媒体类型的查询结果。

4. 视频信息的内容标引

随着计算机和网络技术的快速发展、信息高速公路的建设,以及多媒体的推广应用,视频资料迅速增长,并随之出现了视频数据库、数字电视、视频点播、远程教育、远程医疗等新的服务形式和信息交流手段。如何有效地组织、处理和检索视频信息,已成为计算机应用和信息管理领域的热点课题。

由于基于文本的视频信息标引有不能真正反映视频信息的内容等局限性,基于内容的视频信息分析就成为标引视频信息的选择。基于内容的视频分析从提出到现在,大致经历过三个阶段。第一阶段的研究主要集中在视频结构的分析和浏览上,其中涉及的关键技术有:镜头边界检测、关键帧提取和场景的合并。第二阶段的研究主要围绕相似性检索展开,核心问题是特征的提取和特征空间距离的度量。前两个阶段的研究在一定程度上解决了视频检索和浏览的问题,并对视频管理和获取技术进行了有价值的探索。第三阶段的工作提出了面向语义的信息提取,这也是目前基于内容视频分析研究的热点,该阶段工作的根本目标是构建从底层特征到高层语义之间的桥梁,最终形成符合人类思维形式的信息索引和检索方式。

当前,部分视频检索系统已经通过利用视频图像的基本特征,以及综合各种视觉特征进行联合检索,实现了基于内容的标引与检索,但由于人对视频内容的理解是建立在人类已有知识基础之上的,而这些低级特征无法反映经验知识。如何定义及理解视频包含的语义信息;计算机如何自动提取视频的语义信息,使其尽可能与人对视频内容的理解保持一致,使计算机检索视频的能力接近人的理解水平,从而实现视频检索语义标引与检索,即基于语义的视频标引与检索技术;等等;这些应该成为视频信息标引的研究方向。



本章小结

虽然不同类型分类法的类目体系和标记系统具有自身的特点,其使用规则和方法也存在差异,即使是同一类型的分类法,在上述方面也不尽相同。但是,分类法原理的有关内容是


各种分类法编制和使用的共同基础。在运用特定分类法对信息资源进行分类标引时，只有严格遵循其相关规则和方法，才能保证分类标引的质量，充分发挥分类法在信息整序中的重要作用。

主题法是信息组织的基本方法之一，它直接以表达信息资源主题内容的语词作为标识，以字顺为主要检索途径，以参照系统揭示词间关系的组织和检索信息资源的方法。信息的自动分类与自动标引具有重要意义。



问题讨论

1. 传统分类法包括哪些基本类型？各自具有哪些特点？
2. 层累标记制与顺序标记制的特点是什么？
3. 分类标引的特点和作用是什么？
4. 分类标引的基本分类规则和一般标引规则的内容各是什么？
5. 主题法的特点及类型有哪些？
6. 主题标引的原则是什么？
7. 关键词语言的类型主要有哪些？
8. 简述自动标引的过程。



第6章


信息排检法

本章引言

信息排检法，全称为信息编排检索方法，是指将各种信息有序存储在信息系统、方便人们检索所需信息的编排方法。编排的目的是为了检索，检索则需要按编排的规则去进行。排与检是前后连接、密不可分。因此，信息排检法从信息组织角度看是“编排法”，而从信息检索角度看是“检索法”，所以统称为信息排检法。

目前比较常用的方法有字序排检法（包括音序排检法和形序排检法）、分类排检法、主题排检法、时序排检法、地序排检法及其他排检法。如果说信息著录和信息标引是对信息描述生成信息的特征标识，那么信息排检就是对这些标识排序，组成有序化的信息集合，以便于获取所需要的信息。

本章重点

- 音序排检法；
 - 部首排检法；
 - 分类排检法；
 - 主题排检法；
 - 其他排检法。
- 

6.1 字序法

字序法（字序排检法）又称字顺排检法（字顺法）或查字法，是按一定的顺序排检单字或复词的一种方法。本质上是利用事物名称字顺排检信息的方法。现代中英文字典常用的排检法就是字序法（汉语字、词典或以音序为主，或以部首为主，或以笔画笔顺为主，或以号码为主），再采用其他方法作为辅助。在信息检索系统中，信息资源题名（书名、刊名、篇名、网站名、网页名）、著者名称等检索信息资源的检索标识一般都是按字顺排检的。社会生活中的人名录、机构名录、产品目录、货物清单等也都常用这种方法。字序法的特点是排检款目依字而序，比较直接和直观，便于将字面形式相同或相近的条目集中一处。同时也要看到，采用字序法组织信息时，排序结果只能显示表达信息概念的语词符号在音、形方面的联系和差异，很少或基本上不能反映信息内容上的联系。

既然字序法是建立在信息资源所用文字基础之上的，那么文字就有中外文字之分，汉字则有字音与字形之别。

在联合国最近发表的《2005 年世界主要语种、分布和应用力调查报告》上，汉语被排在第二位，仅次于英语，排在德语、法语、俄语、西班牙语、日语之前^①。汉字是记录汉语的书写符号系统，有着几千年的历史。从甲骨文、金文到小篆，汉字都没有固定的笔画和笔顺，古文字阶段的文字可以说都是由线条构成的，直到隶书阶段才从线条变成笔画。汉字不同于拼音文字，拼音文字是由字母按线性排列的一维组合，而汉字是由笔画、偏旁等构字部件组合而成的，这些构字部件在横向和纵向两个坐标上以上下、左右、内外等方式组合成二维平面图形，本身缺乏完备的表音系统，因此建立字序比字母文字复杂得多。所以字顺排法也是多种多样的。归纳起来，中文常用的字序法包括音序法（汉语拼音排检法、注音字母排检法等）和形序法（部首法、笔画与笔顺法、四角号码法等）。

音序检字法和部首检字法是汉字检字法中最受欢迎的两种检字法。它们已经广泛地应用于现代汉语字典等工具书的编纂，成为中国当代两大主流检字法。音序检字法是后起的先进检字法。它产生后就在汉语工具书中取代部首检字法，成为第一检字法（核心检字法）。

外文字序排检法依据其语言字母的固定次序排序。

6.1.1 汉字音序排检法

汉字音序排检法是将信息款目按照汉字的读音及表示读音的字母顺序排列款目的方法，即按汉字字音顺序编排汉字的方法。这种排检法又分为以下几种：汉语拼音字母排检法、注音字母排检法、韵部排检法、威妥玛式拼音排检法及声部排检法等，其中汉语拼音字母排检法最为常用。

由于汉字是表意文字，音序排检法的局限性就比较大，查字的人必须懂得要查的字的读音，不知道或搞错了读音就无法找到要查找的对象。人们之所以查字典，通常是因为不知道某个字的读音才去查字典的，这时怎能用音序法去查？虽然现代汉字中形声字占的比重较大，但是由于三千多年来汉语语音发生了很大的变化，就造成了形声字的老化——表音功能的退化。因此，音序排检法的优点是单字排序简单准确，查字速度快捷，缺点是不知字的读音就无法利用此检字法，不能独立完成检字任务。因此，汉字音序检字法是一种不完全的半独立的检字法，不能等同于外语词典的音序检字法^②。

① 世界主要语种排行。http://www.fubusi.com（2006-3-27 9:01:00 更新，2009-08-08 查询）。

② 李学金。从外语词典视角看汉字检字法的统一问题。广西广播电视大学学报，2005，15（2）：50～54。

1. 汉语拼音字母排检法

汉语拼音排检法,又叫汉语拼音检字法,简称拼音法,是一种依照汉字的汉语拼音字母顺序来编排词目的排检方法。

汉语拼音字母排检法是目前中文信息排序使用最广泛的一种方法。此法依据我国 1958 年 2 月 11 日公布的《汉语拼音方案》,采用 26 个拉丁字母作为汉语拼音,排列次序依照国际惯例,从 A 到 Z。在 ABC…Z 的 26 个字母中,除“l”、“u”、“v”三个字母不做音节头外,其他 23 个字母都可以做字头,共分 23 部。其排检规则是先按汉字拼音的第一个字母顺序排;第一个字母相同时,再按第二个字母顺序排,依此类推;字音相同时,再按声调的阴平(ˉ)、阳平(ˊ)、上声(ˇ)、去声(ˋ)及轻声顺次排列。例如,啊(a)、爱(ai)、把(ba)、国(guo)、来(lai)、们(men)、我(wo)、祖(zu);又如,巴(ba)、拔(ba)、把(ba)、爸(ba)。若字节和声调都相同,则按汉字的笔画笔形排列,如禾、合、纥、何……。汉语拼音字母排列法也可以用汉字所组成词语的音序进行排列。例如, chun feng (春风)、chun guang ming mei (春光明媚)、chun ji (春季)、chun jie (春节)、chun yi ang ran (春意盎然)。当然,由软件程序驱动的计算机、手机等,其文字处理系统在使用汉语拼音输入法输出的同音字并不是完全依笔形笔画排列的,而是按照其字库的已有设定或使用频率排列的。

《汉语拼音方案》草案是在 1957 年 11 月 1 日国务院全体会议第 60 次会议上通过的,并于 1958 年 2 月 11 日第一届全国人民代表大会第五次会议批准推行。该方案由字母表、声母表、韵母表、声调符号、隔音符号五部分组成。自此,按汉语拼音排检工具书成为一种主要的方法。1982 年国际标准化组织承认汉语拼音为拼写汉字的国际标准,汉语拼音开始走向世界。

我国现行的一般语文性字典、词典等工具书,大多采用汉语拼音法,如《新华字典》、《现代汉语词典》、《汉语主题词表》及《中国分类主题词表》(第二卷)等都采取这一排检法。

汉语拼音字母排检法的主要优点有如下几个。

(1) 方法简单、易于掌握。作为信息编排方法,任何具有汉语拼音和汉语语言文字基础的中、外读者可以很快熟悉和掌握。

(2) 查找文献快速、准确。读者只需知道读音,就可以直接找到相关信息。

(3) 比较科学,符合国际上工具书按字顺编排的习惯,也适用于计算机检索。

汉语拼音字母排检法的主要缺点是要求使用者必须学会普通话,发音要求准确等。如果不知字的读音,或者读音不准,就无法使用汉语拼音字母排检法查检。其次,汉语同音字很多,对同音字的编排方法,不同的信息系统也有所不同,有的按笔画多少排序,有的按部首排序,这是利用汉语拼音字母排检法需要注意的问题。此外,在有的方言中,声母 l、n 不分,韵母前鼻音(如 en)、后鼻音(如 eng)不辨,特别是对南方人来说,大量的声母混淆在一起,用这种音序法检字也是困难的。因此用这种方法编排的工具书,都附有其他索引以备查检。

需要指出的是,在运用汉语拼音字母排检法的字顺检索系统中,检索标识中可能不仅有汉字,还会出现阿拉伯数字、罗马数字、拉丁字母甚至标点符号等其他符号,不同的信息系统对此处理可能不同,这需要检索者关注一下该检索系统的相关详细排序规定与说明。

2. 注音字母排检法

注音字母排检法又称注音字母音序法。注音字母音序法中的注音字母是以北京语音为标准的一套字母。它是《汉语拼音方案》公布实施之前,为汉字注音和推广普通话而设计的一套音标。注音字母(参见《现代汉语词典》附录)是由汉字偏旁改造而成的 40 个字母,其中含声母 24 个、韵母 16 个。目前仍使用的有 37 个(声母 21 个,韵母 16 个)。它的特点是字母全部都是笔画简单的古汉字(楷书的篆文古体字)。音节的拼写采用声、介、韵 3 拼法,

声调则另加标记。它的排列顺序是先声母,后韵母,韵母的 l、ㄨ、ㄩ 依 1931 年公布的国音字母表列于儿韵之后。声母从 ㄅ(b)、ㄆ(p)、ㄇ(m)、ㄈ(f)、ㄉ(d)、ㄊ(t)、ㄋ(n)、ㄌ(l)、ㄍ(g)、ㄎ(k)、ㄏ(h)、ㄐ(j)、ㄑ(q)、ㄒ(x)、ㄓ(zh)、ㄔ(ch)、ㄕ(sh)、ㄖ(r)、ㄗ(z)、ㄘ(c)、ㄙ(s)。韵母从 ㄚ(a)、ㄛ(o)、ㄜ(e)、ㄝ(ie)、ㄞ(ai)、ㄟ(ei)、ㄠ(ao)、ㄡ(ou)、ㄢ(an)、ㄣ(en)、ㄤ(ang)、ㄥ(eng)、ㄦ(er)、ㄨ(i)、ㄩ(u)、ㄩ(u)。声母与韵母拼音时,先声母、后韵母并按音序先后排列,如“八”(ㄅㄚˇ)、“拨”(ㄅㄛ)、“爬”(ㄆㄚˊ)、“坡”(ㄆㄛ)。同声同韵的字母再按阴平、阳平、上声、去声四声排列,如 ㄅㄚˊ(哩)、ㄅㄚˊ(离)、ㄅㄚˊ(李)、ㄅㄚˊ(立)。

注音字母是中国历史上第一套法定的拼音字母,它的公布和使用是汉字注音走向拼音化的开端。1913 年 2 月 15 日,注音字母由中国读音统一会制订,于 1918 年由北洋政府教育部公布。1930 年国民党政府改名为“注音符号”。1958 年《汉语拼音方案》公布前,注音字母是中国 40 多年中一直通行的汉字注音符号,它对于统一汉字读音、推广普通话、普及拼音知识起着重要的作用。1958 年前后出版的字典(词)典大多数都以注音字母注明其读音,如《汉语词典》、《新华字典》(1956 年、1959 年版)和《同音字典》等。目前我国港台地区出版的某些工具书及计算机、手机信息存储交流的汉字输入、输出法仍有采用。

注音符号在当时的人文背景中采用的是传统的民族形式的字母,这从文字发展的角度看是不明智的。它的缺点是在标注汉字时,与汉字的区别不大,没有采用音素拼写法,符号本身也没有彻底音素化,其中有好几个韵母还可以进一步分解为更小的语音单位,字母拼写与实际读音仍未切合,没有采用国际通行的拉丁字母方便。因此,现代只有《新华字典》、《四角号码新词典》、《现代汉语词典》等少数辞书在汉字词目后面有附注而已。今天的“汉语拼音方案”中的字母表、声母表和韵母表对应的符号就是注音字母。

3. 韵部排检法

韵部排检法也称“声韵法”,是我国古代按音韵排列汉字的一种方法。汉字字音都是单音节,每个音节都由声、韵、调组成。把韵母相同的字集中排列在一起就构成一个韵部。韵部顺序大体可分为三种。第一种,按汉字平、上、去、入四声分类(其实就是古汉语的四个音调,和现在汉语的声调不同),同一声调下再分韵部,韵部之内再按同声字分类编排,如《广韵》(宋陈彭年等修订)。现存韵书大都按这种方法编排,即“以声为纲,以韵目为经”分列韵部。第二种,先分韵部,韵内分声调,声调内再按同声字分类编排,如《中原音韵》(元周德清著)。第三种,先分韵部,韵内分同声字,同声字内再按四声分别编排,如《韵略易通》(明兰茂著)。

按韵部编排的字典称为“韵书”。我国在不同的历史时期有不同的韵部。中国古代影响较大的韵部系统有《广韵》和《诗韵》。宋代《广韵》分 206 个韵部,为《集韵》所沿用。南宋的《礼部韵略》并为 107 部,金元又改为 106 部,称《平水韵》,明代《洪武正韵》为 76 部,一般认为 106 韵部《平水韵》较为通行。自宋以来流行的平水韵(《平水新刊礼部韵略》)将其删并为 106 韵,从而成为文人作诗用韵的规范。“声韵法”本来是为了科举考试做诗用的,但是它的产生却深深地影响了古代工具书的编排检索方法。一些主要工具书也按平水韵编排,如清代编的《佩文韵府》、《经籍纂诂》及现代编的《辞通》等都按词目的末字分韵编排。

韵部排检法先按声分部,然后在每部之下列若干韵。古代文人作诗讲究“平”“仄”两音。其实“平”是指平声,“仄”指古汉语中的上、去、入三声。现代汉语中,入声消失,平声分“阴平”“阳平”,对应现代汉语拼音的四种音调,即:阴平(第一声);阳平(第二声);上声(第三声);去声(第四声)。一个韵部包括很多字,先用数字表明属于第几韵,

是为了古体诗中换韵方便；再用一个字代表这个韵。例如，“八齐”就是指第八韵部“齐”韵，“归”和“微”都属于“五微”韵（参照下文的《平水韵》）。

《平水韵》106部韵如下。

上平声：1东、2冬、3江、4支、5微、6鱼、7虞、8齐、9佳、10灰、11真、12文、13元、14寒、15删。

下平声：1先、2箫、3肴、4豪、5歌、6麻、7阳、8庚、9青、10蒸、11尤、12侵、13覃、14盐、15咸。

（注：由于平声字多，沿用《广韵》将平声字分上下两卷，上平声就是指平声上卷，下平声就是指平声下卷。）

上声：1董、2肿、3讲、4纸、5尾、6语、7麌、8荠、9蟹、10贿、11轸、12吻、13阮、14旱、15潜、16铣、17篠、18巧、19皓、20哿、21马、22养、23梗、24迥、25有、26寝、27感、28俭、29赚。

去声：1送、2宋、3绛、4寘、5未、6御、7遇、8霁、9泰、10卦、11队、12震、13问、14愿、15翰、16谏、17霰、18啸、19效、20号、21箇、22馊、23漾、24敬、25径、26宥、27沁、28勘、29艳、30陷。

入声：1屋、2沃、3觉、4质、5物、6月、7局、8黠、9屑、10药、11陌、12锡、13职、14缉、15合、16叶、17洽。

按韵部排列的工具书，主要用于查检古代的字书、韵书、类书等。要查找古代资料，必须会用这种检索方法。用此法编的工具书有《永乐大典》、《佩文韵府》、《辞通》、《经籍纂诂》、《九史同姓名略》（按条目首字分韵编排）等。检字时，先确定所要查的字属何声，然后确定何韵，在韵下再找出所要查到的字。使用者若不熟悉某一汉字的韵部，可先查《中华大字典》、旧《辞海》等工具书。还可借助新印本所附的索引先查出该字的韵部，再按韵部去查。如新印本《佩文韵府》和《辞通》都编有首字四角号码和笔画索引。《平水韵》在古代简便科学，但烦琐不便记忆，到近代已被淘汰。

4. 威妥玛式拼音排检法

英国人威妥玛（Thomas F. Wade, 1818—1895）于1867年编写的汉语课本《语言自述集》中，设计了用拉丁字母拼写汉字的较完善的方法，称为威妥玛式拼音法（Wade System），后来经过改进，又被称为威妥玛—翟理斯式（Wade-Giles System），是英文中拼写汉字的主要方法，通行已久。编排工具书时，依拉丁字母的次序排列，不分声调。哈佛燕京学社编辑出版的引得所附拼音检字即采用此法。西方编纂的汉学工具书也大都用此种编排方法。过去用威妥玛式拼音法拼写中国的人名、地名，现已统一改用汉语拼音字母拼写。

要注意威妥玛式拼写方法与汉语拼音字母的区别。如 ch'un（春）、hsia（夏）、ch'iu（秋）、tung（冬）、jih（日）、yueh（月）。此外，威妥玛式不仅 zh 与 j 不分，而且往往不标送气符号，所以，“朱、楚、居、瞿”可能都拼做 chu^①。

5. 声部排检法

声部排检法是将汉字按古声母分类排列汉字的方法。现在沿用的是唐宋时古人归纳出的36个字母。这36个字母是：见、溪、群、疑、端、透、定、泥、知、彻、澄、娘、帮、滂、并、明、非、敷、奉、微、精、清、从、心、邪、照、穿、床、审、禅、影、晓、匣、喻、来、日。按古声母法编排的词典性著作有清代王引之编著的《经传释词》等。

① 叶继元. 信息检索导论（第2版）. 北京：电子工业出版社，2009.

6.1.2 汉字形序排检法

汉字的形体结构具有某些共同的特点，加以归纳分类，寻求规律，依次排列，即成为检索汉字的一种主要方法。查字时，从汉字的字形去检索其读音和字义，符合人们从形出发求音求义的查字要求。形序排检法就是根据汉字的形体结构特征及书写方式，依次编排和查检汉字的方法。常用的形序法有部首排检法、笔画笔形法和四角号码法等。形序排检法要求对字的部首查找准确，对字的笔形、写法掌握准确。

1. 部首排检法

部首排检法又称部首查字法，是根据汉字部首偏旁的结构特点排检汉字的方法，简称部首法。它是我国工具书的传统排检法，是根据汉字的形体结构特点，利用其偏旁（汉字的各个组成部分）的同一性来编排条目的方法。汉字的形体结构，除少数属独体字外，大多是合体字，即由形旁（也称义符）和声旁（也称声符）组成的形声字，彼此之间具有一部分相同的形体，把这些形体相同的部分归为一类，称为偏旁或部首。在独体字中，有的本身就是部首。部首按笔画数多少排序。笔画数相同的部首，依起笔笔形排序。同属一个部首的字，先按笔画数（一般不包括部首的笔画数）排序，笔画数相同者，再依起笔笔形排序。把部首按顺序加以排列检字，就是部首排检法。它是我国工具书中最普通的一种排检方法。从东汉许慎的《说文解字》到《康熙字典》、《中华大字典》、《辞海》等都采用了这种方法。

1) 部首法的传承

在我国历史上，最早出现的检字法就是部首检字法。它是由东汉文字学家许慎在《说文解字》里首创的。他总结出“以部相从”（亦即“据形归部”）的收字准则和检字法，把 9543 个字分“部”归入 540 个部首，具有同一偏旁的字归为一部，同部字的共同偏旁排列在一部之首。故称“部首”。部首和部内字的排列，都按照笔画的多少，由少到多，编排成书。许慎之前的字书，可以分为两大类：第一类是按照义序法排列的只解释字义的字书，可以简称为义书，如《尔雅》、《方言》、《小尔雅》等；第二类是只编纂文字和只解释文字的字书，也就是通常所说的童蒙识字课本，如《史糟篇》、《仓颌篇》、《爱历篇》、《博学篇》、《凡将篇》、《急就篇》、《元尚篇》、《训纂篇》、《傍喜篇》等。义书把同义词排在一起，童蒙识字课本大都是三言或七言一句（也有四言一句的），隔句押韵，每一章一韵到底。因此，这两种字书的字词排检没有规律，从严格意义上说，不能算是说解文字的字典。

许慎研制的这种部首法使纷繁芜杂的汉字变得有序可循，便于检索学习，沿用至今。需要注意的是，许慎确立的 540 个部首基本上都是义符，即形旁。初期的部首法也存在一些弊端：何者为“部”，难以认定；部首数量过多，不便翻找检查；规则较多则乱，难以形成规范。南北朝梁顾野王著《玉篇》立 542 部，辽僧行均著《龙完手鉴》立 242 部，明梅膺柞著《字汇》合并为 214 部。时至公元 1710 年（清康熙四十九年）朝廷命张玉书主编《康熙字典》，呈现诸多改进。根据楷书字形分部和建立部首，分部标准以字形为主，部首的排列和每字所属次序都突破了《说文解字》以意义为依据的原则，改为以字形笔画少多为序。特别是压缩了部首数量，由 540 个减少为 214 个，大大便利了检字和学习。

1953 年出版的《新华字典》更是精简为 189 个部首。近年我国颁定《汉字统一部首表》规定部首数为 201 个。现行辞书中应用较多的有 214 部、250 部、200 部、189 部四种。《辞源》用 214 部，新《辞海》用 250 部，《汉语大词典》和《汉语大字典》均用 200 部，《新华字典》和《现代汉语词典》检索用 189 部。214 部是自《字汇》、《康熙字典》以来的传统部首，250 部参照了 1964 年汉字查字法整理工作组的推荐方案，200 部参照了 1983 年汉字部首排检工作组拟定的 201 部，189 部依据新华辞书社拟制的部首检字法。2004 年出版的第 10

版《新华字典》在 201 个部首的基础上另加 79 个“附形部首”，读者所面临的部首总数是 280 个，这自然是形势所需。

时代的演进和社会的需要迫使部首法研究有新的突破。“新部首检字法”呈现于 1988 年出版的《新部首大字典》中，定部首 57 个，附变体 101 个。据方法发明人亦字典编著者在前言中简介和序言撰写人推荐，这一方法的优点是选取最常用字、最常用笔画、偏旁作为部首并当“字母”用；按字形结构特点合理分割拼接，实现汉字“字母化”，做到了易学、易记、易用，并且做到 5 万多字基本上无重码的程度，为实现自动化检索创造了条件。

从部首法的演变可知，由于成书年代不同及其他一些原因，各类工具书采用的部首并不完全相同，读者查阅这些书时，需要了解部首设置情况。

2) 部首法的使用

部首之建立，揭示了汉字形符在字形演变过程中的重要地位，发现了“据形求义”的科学方法。因此，540 部归纳了汉字构形的基本符号，抽象出汉字表义的基本部件。部首编序法从东汉许慎创制至今，将近两千年来，在字典、辞书及其他中文工具书的编纂中占有极为重要的地位，功不可没^①。

部首法编排工具书，就是按照汉字楷书字体，对工具书中所收的汉字进行分析，归并成若干部，以部首的笔画多少依次排列。同一部首中的字，再按照其余（除部首以外的）笔画多少作为编排次序。

汉字可以分为独体字和合体字。独体字是没有表音成分的纯粹表意字，数量很少。绝大多数汉字是合体字，即由形旁（也称意符）和声旁（也称声符）组成的形声字，如“搬”、“拦”、“指”等字，其中“扌”是形旁，“般、兰、旨”分别是声旁。

按照字形结构，把含有相同偏旁的汉字归为一类，这种偏旁就是部首。如“赴”、“赵”、“起”、“超”、“趣”等都有相同的偏旁“走”，属“走”这一部首。把同一部首的字，归并为一类，同类部首的字，再按笔画多少顺序排列，这就是部首检字法。

使用部首法必须首先分析字形结构，掌握选取部首的一般原则和特殊规定，要注意一些部首与该部所收字的偏旁写法不同，如牛部：“牡”、“牝”、“物”等。

现以《辞海》为例，说明部首检字法的规定：一般优先采用字的上、下、左、右、外五个部位的偏旁作为部首；其次是中间和左上角。如果一个字具有几个部首，则按以下规定的次序选取。

- (1) 上下都有部首的，取上不取下。例如，“贡”字查“工”不查“贝”。
- (2) 左右都有部首的，取左不取右。例如，“柏”字查“木”不查“白”。
- (3) 内外都有部首的，取外不取内。例如，“闪”字查“门”不查“人”。
- (4) 在同一部位有多笔和少笔的，取多笔不取少笔。如，“章”取“音”，“空”取“穴”。
- (5) 部首位置无从选取或所在位置不符合规定的，按起笔取单笔部首或按笔画排入难字表中。起笔顺序一般为“一”（横）、“丨”（竖）、“丿”（撇）、“丶”（点）、“乛”（横钩）。

部首法一般以笔画总数从少到多进行排列，笔画数目相同的又按一、丨、丿、丶、乛五种笔形顺序排列。相同笔形的再按汉字笔顺结构排列，一般按上下、左右、内外、中间两边、其他的笔顺顺序。

在使用部首法查阅工具书时，应首先分析字形结构，搞清部从，查出部首，数清部首以外的笔画数目，然后查字。对于一些难以确定部首的字，可查“难检字表”或辅助索引。

部首法适应了汉字的表意性特点和结构特点，把字形复杂、数量庞大的汉字划归入一二百个部首里，符合人们由字形求音、解义的习惯，对不认识的字或读不准普通话语音的字，

^① 燕荣晖. 试论部首编序法的发展. 江汉大学学报, 2000, 17 (4): 130~134.

都可以利用部首法查找出来。因而《汉语大词典》、《辞源》(修订本)、《辞海》(1979年版、1989年版、1999年版)等仍然以部首法来编排字目,同时辅以其他排检方法。部首法的缺点是:由于汉字形体极其复杂,部首位置又不固定,规则多且不统一,所以有些字的部首很难确定,以至于查检较难。

2. 笔画笔形法

笔画是汉字的基本构件,笔画构成字形,汉字由笔画构成。笔画笔形法是按照笔画数目和起笔(首笔)笔形的先后次序排检汉字的方法。它有三种应用形式。第一种是先按笔画多少来归并汉字,笔画少的在前,多的在后;笔画数相同者,再依起笔笔形排序。笔画和笔形均相同的字,则依其字形结构排序。第二种形式是先按笔画多少来归并汉字,笔画数相同的,再依部首归类排列先后顺序。第三种形式是先按笔形分类,统一笔形后再按笔画顺序排列,这样的排检方法也叫笔顺法。

汉字的基本笔形是点(丶)、横(一)、竖(丨)、撇(丿)、捺(㇏)5种。两种或两种以上的笔形连用又组成复杂的折笔(如“冫”等)。汉字书写时起笔只用丶、一、丨、丿、㇏(包括竖折、撇折)5种笔形。

按笔画多少排序的优点是排检原理简单,易学易检,使用方便,只要识字就会使用。掌握笔画法的前提是要能辨清一个字的笔画。由于汉字笔顺复杂,汉字的繁体字和简化字的笔画不同;印刷体和手写体、新旧字形笔画往往也不同。因此,其缺点是很多字的笔画不容易数准确,同时,同笔画的字太多(如《康熙字典》中12画的字就有3642个),检索速度很难提高。采用笔画法编排的工具书较多,有《中国人名大辞典》、《中国古今地名大辞典》、《十三经索引》等。笔画法还作为辅助检索方法使用。

笔画相同的字,或者按部首归类,或者按笔形顺序排列。由此还要注意起笔的顺序,才能确定起笔的笔形。汉字的笔形书写习惯通常是,先上后下(如“安”),先左后右(如“好”),先外后内(如“同”),先中间后两边(如“水”),先横后竖(如“十”),先撇后捺(如“入”),点在左上先写(如“为”),点在右上后写(如“式”),中间点后写(如“梅”),竖折底后写(如“巨”)。文化部、文字改革委员会于1964年12月联合公布的《印刷通用汉字字形表》(文字改革出版社1986年版)规定了6196个字的字形和笔顺,可以作为确定笔顺的规范。

利用笔形顺序排检,有的只用第一笔的笔形(起笔笔形法),有的利用各笔的顺序排列,而且笔顺也不统一,如有3种顺序的4种笔形(丶、一、丨、丿;丶、丨、丿、一;一、丶、丿、丨),5种笔形(一、丨、丿、丶、㇏;一、丨、丶、丿、㇏),还有7种笔形的。可见,笔顺法虽然简单,但书写习惯不同,笔顺和起笔有时很难确定,就是在现行的一些工具书中,某些字的笔顺也有分歧。因此只有少数几种工具书利用笔顺法编排,如姚德芸编的《古今人物别名索引》。现在主要用起笔笔形作为笔画法的补充。

笔画笔形法的优点是道理简单,限制条件少。由于汉字结构复杂和字数众多等条件的制约,还存在不易看、数不准的缺点。

3. 号码法

号码法是根据汉字的笔形结构进行编码排列,然后根据号码顺序检字的方法。号码法是形序排检法的一种变形,它把汉字分解为若干种笔形,每种笔形用数字作为代码,然后把每个字的笔形代码连接为号码,再按号码大小排列的一种检字方法。这种方法的优点是只要记住笔形代码、号码的位置次序,则数字简单明了,检索迅速,使用便利。其缺点是学习和掌握比较困难,只有经过反复练习,才能运用自如。在常用的工具书中,主要有四角号码法、中国字庖丁法;在计算机信息存储系统中有五笔字型法等。

1) 四角号码法

四角号码排检法(四角号码法),又称四角号码查字法,由原上海商务印书馆王云五先生(1888—1979)创制。四角号码法是根据汉字方块形体的特点,用数字来描述汉字四角的笔形,使每个汉字拥有一组数码,再按号码的顺序编排汉字先后次序的排检方法。它将汉字四个角上的笔形归为10类,并分别用0~9共10个数字代表。

笔形名称:头 横 竖 点 叉 插 方 角 八 小。

对应号码:0 1 2 3 4 5 6 7 8 9。

这10种笔形取号有一个帮助记忆的歌谣:横一竖二三点捺,叉四插五方框六,七角八八九是小,点下有横是零头。当然,10种笔形名称中每种一般包括几种相近笔形,如横(1)中包括一(横)及其变形“乙”,八(8)中包括八、人、入等笔形。因此,使用时注意对照工具书中编排体例中的《四角号码笔形代码表》。

四角号码法是对每个汉字都取四角,即在字的左上、右上、左下、右下四个角取笔形,把四角的笔形用数字表示出来,按上述取角的顺序(先上后下、先左后右)连接起来,就成为四角号码了。由于汉字数量大,四角号码法为了避免重码字过多,有时需要再依据四角号码从汉字的右下角笔形再取一个附号。如“玉”字完整的四角号码是10103,其中“3”是附号。四角号码按以下顺序排列汉字:按号码大小由小号到大号排列;同号码的字按附号顺序排列;四角和附号均相同的字,再按各字所含横笔的数目顺序排列;条目首字相同的,再以条目第二字的头两角的号码大小为序。如,“物”字取号为2752,“盛”字取号为5310,如此翻至相应的页码,就可检得所需汉字了。

四角号码法原称“四角号码查字法”,最早发表于1926年3月出版的《东方杂志》,两年后又出版了“第二次改订四角号码检索法”,两次均以王云五名义发表。该检字法经中国文字改革委员会汉字查字法整理工作组的若干修改,公布了新的“四角号码查字法”(草案)。1977年出版的《四角号码新词典》(修订重排本)即以新法编排,使四角号码法有新法、旧法之分。

“四角号码检字法”最早用于商务印书馆1928年出版的《四角号码学生字典》,以后得到普遍应用。推行简化字后,对四角号码作了修改,改称“四角号码查字法”,一般称为新四角号码法。目前出版的工具书,采用新法或旧法的都有。旧法笔形依通行的手写体,而新法笔形则以《印刷通用汉字字形表》的规定为准。新法与旧法取角方法有所不同。《现代汉语词典》和《四角号码新字典》附有《新旧四角号码对照表》,列出了主要的不同点,可以参看^①。

四角号码法曾在多种文史工具书中采用。它是一种独立的排检方法,不必与其他方法结合使用。在使用四角号码编排正文的工具书中,一般都编有使用说明。《四角号码新词典》即是典型一例。有些工具书用四角号码法编排正文或编制辅助索引,如《中国丛书综录》、《二十四史纪传人名索引》、《本草纲目索引》等。

四角号码法的优点是可以见字知码,查检迅速,较为简捷,不知读音也可查到,是一种独立使用的排检法。缺点是取号规则过于繁复,变形的笔形较多,不易掌握,汉字的繁简体及印刷体与手写体还有所区别,对笔形稍有误解即难查到。

2) 中国字度撇法

中国字度撇法中的“度撇”(gǔ xié)二字的意思是放入、取出。它是根据汉字的形体结构,把字形和笔形变成数码的一种排检法。1931年由哈佛—燕京学社引得编纂处负责人洪业所创^②。

① 冯惠玲,王立清.信息检索教程.中国人民大学出版社,2004:116~122.

② <http://www.historychina.net/cns/QSYJ/XLMB/XZSL/honyzs/06/22/2004/9417.html> (2009-8-14 查询).

具体说来,它是将所有汉字分为“中、国、字、度、撇”等五种字体;且“度”之义为“放入”,“撇”之义为“取出”,全称即为“中国字编入检出”之意,与检字法名实相符,且有助于学习使用者记忆。中国字度撇法属于号码排检法的检字体系,先将汉字按字体分为五类,以“中、国、字、度、撇”五字统率,再以罗马数字“I、II、III、IV、V”为定体号码;度撇二字的笔画进行分解后,可得十种部件,分别对应0~9号,然后依汉字四角笔画取4个号码;最后据该字所含方格数确定一号码,于是每一汉字的号码由6个数字组成。此法编排是首先按五种字体分类,然后在每种字体中再按这个字的号码从小到大排序。中国字度撇法同四角号码法相比,重码字少,但取号规则既烦琐复杂,又不太科学,难以掌握,因而使用不广,已被淘汰。哈佛—燕京学社运用它编了64种引得,多为古籍整理类的线索性工具书,但也附上了笔画法、四角号码法和汉语拼音法。如果确实有需要了解和学习,请自行参看邓宗荣著《社科中文工具书使用》,辽海出版社2006年7月第2版第20页。

3) 五笔字型法

它首先是供计算机信息处理用的汉字编码法。五笔字型是由查字法发展而来的一种汉字输入法。它根据汉字的字形结构,从中选定130个部首作为字根,加以分类、编码,并将其排在25个英文键位上。通过字根的组合,可以打出汉字或词组,从而达到见字知码、操作方便、快速输入的目的。五笔字型输入技术已得到广泛的推广和使用。五笔字型输入法是王永民于1983年8月发明的一种汉字输入法。之所以称为五笔,是将汉字笔画分为横、竖、撇、捺、折五种。

汉字编码的方案很多,但基本依据都是汉字的读音和字形两种属性。五笔字型完全依据笔画和字形特征对汉字进行编码,是典型的形码输入法。五笔字型输入法在使用简体中文的地区较广泛,是这些地区最常用的形码输入法。

由于五笔字型编码方案中将汉字众多笔画进行了归类,并分别与键盘上的英文字母键一一对应。这实际上是将汉字与英文字母建立了联系,使人们可以利用国际的通用字母方案将汉字系统排序。

其他号码法还有起笔笔形号码法、六位笔形号码法、三角号码法、五码查字法等,这些号码法尚未推广,只有少数几种工具书使用。随着计算机检索的推广使用,供信息处理用的汉字编码法有很大发展,各种编码方案纷纷问世,仅笔形编码方案就有300多种,目前五笔字型法最为通用,此外区位码、首尾码、声数码等也较为常用,此处不再赘述。汉字编码法的应用和普及,必将进一步促进字顺排检法的改革和完善。

6.1.3 外文字顺法

目前世界上有2000多种语言,多数有文字,适用范围超过5000万人的语言有13种,其中联合国正式的工作语言有:汉、英、俄、西班牙、法、阿拉伯语6种语言。不同的文字有不同的字顺排检法。例如,西文字顺排检法,它以拉丁字母顺序排检西方文字。字母顺序排检法最为常用。外语词典里的单词都是按字母表字母顺序排列的,这在全世界都是统一的。各种语言的字母数量不尽相同,英语词典和法语词典按26个字母排序,西班牙语词典按28个字母排序,俄语词典按33个字母排序,阿拉伯语词典按28个字母排序,韩国语词典按40个字母排序,吸收了许多汉字的日语词典按《五十音图》排序。外语词典编纂经验告诉我们:一种语言词典只有一种单词固定排序的方法,而这种单词固定排序是由该语言字母的固定排序所决定的。

字母顺序排检法就是机械地按字母顺序排列。使用这种方法进行排检,不必预先掌握或记住任何既定的组织体系或排列方法,直接利用已知的或设想适用的检索词即可查到所需信息。

字母顺序排检法有两种不同的应用形式。第一种是逐词排列法,即 **Word by Word**。以参与排检的各个独立的词为排检单位,逐词相比,凡第一个词相同时则比第二个词,第二个词相同时则再比第三个,依此类推。第二种是逐字母排列法,即 **Letter by Letter**。所有参与排列的项目,无论是单词、词组或句子,不管其字数的多少或长短,均视为一个排列单位,按字母逐个相比。如表 6.1 所示^①。

表 6.1 英文字顺法两种排列方式的比较示例

Word by Word	Letter by Letter
Air Conditioning	Air Conditioning
Air Cushion Vehicles	Aircraft
Air Force	Air Cushion Vehicles
Air Pollution	Airfields
Air Transport	Air Force
Aircraft	Air Pollution
Airfields	Airports
Airports	Air Transport

逐字母排列法将标目用词连起来作为一个单元逐个字母进行排比次序,其优点是规则简单,排列方便,但不便于集中相同的单词,破坏字面成族,因而很少被采用。逐词排列法以词为单位,单词内再按字母的先后次序编排,这种排列法有字面成族的效果,能集中所有相同的词,比较符合人们的检索习惯,比逐字母排列法科学。因而在实际应用中,“**Word by Word**”排检法应用较广,国内外大多数图书馆的字顺目录几乎都采用此种排检法,一些历史悠久的著名工具书也多用此法,因为它能把首词相同的款目集中于一处,达到族性检索的作用。例如,在传记词典中,同姓氏的被传者将集中于一处。

西文字顺排列涉及许多具体问题,亦有排列与不排列的选择,一旦选定一种排列方式,则应一以贯之。例如,有些排列法有如下规定。

(1) 其字符排检顺序是空格、破折号、连字符、斜线、圆点、数字(0~9)、英文字母(A~Z)、非罗马字母。所有字母的变音符号及标点符号一律不予考虑。

(2) 符号“&”转换成所代表的相应文种的原连接词词形排列。

(3) 缩写词按其在标目中出现的书写形式排列。

(4) 首冠词不予排列。但作为地名、人名整体不可缺少的首冠词和作为数词的不定冠词要排列;非英文文种的复合冠词应予以排列。

(5) 地名和人名的前缀作为一个单词排列,但前缀与地名、人名相连或其间加一省略符号连接者则与其后的词作为一个整词排列。

(6) 年、月、日依其出现的形式(文字或数字)排列。

(7) 国名简称,按其全称排列。

(8) 会议召开的年代、届次,手册或年鉴的年代等,数词一律排在会名、书名后。

(9) 书名中的数词、年代、版次、卷次等,一律按数词的自然顺序排列。

(10) 作为书名组成部分的拉丁字母数词,均按拉丁字母排列,不按阿拉伯数字排列。

此外,按照“**Word by Word**”法来排检,对复合词、带前缀的姓氏及其他特殊语言现象可能会产生混乱的情况,必须辅之具体的排列细则加以限定才能确保排检结果的一致性。因

^① 王作梅,严一桥,孙更新. 西文文献编目. 武汉: 武汉大学出版社, 1997.

此，在西文字顺排检时应注意一系列细节问题，例如：

- (1) 关于在款目之首或当中的虚词（如连词、介词等）的取舍问题；
- (2) 关于不同拼写形式词的排检问题，如 Catalog 与 Catalogue, Labor 和 Labour 等；
- (3) 关于某些姓氏前缀的处理问题，如 M, Me, Mac 和 St, Sant 等；
- (4) 关于同名异义款目的排检问题，如同为 Washington 的一组款目，如人物、地方、事物、事件等，孰前孰后，如何区分；
- (5) 德文中的 4 个变音字母 A、U、O、B，根据国际惯例，可按 A=AE, U=UE, O=OE, B=BE 来排^①。

俄文字顺排检法主要按照 33 个俄文字母的顺序排列。也有两种排检方式，即逐字母排检和逐词排检。

日文字顺排检法有 3 种情况：①按照五十音图的顺序采用平假名排检；②按照五十音图的顺序采用片假名排检；③借用汉字排检法按照部首笔画等方式排检。

6.2 类序法

可以说任何有深度的信息存取系统都要求能够按主题内容及其相互关系进行编排和检索，6.1 节的字序法在排检庞大综合性的信息集合时，突出不足便是使相关信息内容分散开来，不易使人们了解它们之间的内在联系。而类序法可以解决这一问题。类序法（分类排检法）又叫类别法，指按学科属性或事物性质及其概念逻辑关系分门别类组织编排信息的方法。这是一种从概念的系统性与相互关系的角度对信息进行揭示和检索利用的基本方式。

类序法在传统上主要分为两类：一类是按学科体系分类；另一类是按事物性质分类。前者如依据《中国图书馆分类法》对现代文献进行分类形成的分类目录等。后者是按事物属性或知识范畴排列信息，如古代类书、政书等工具书，以天、地、人、事、物为纲。现代的年鉴、手册等工具书通常也按类序法排检。类序法将信息分门别类加以集中，充分展现信息间的逻辑关系。类序法不仅用于分类目录、分类索引、类书等，社会活动的各个领域也是比比皆是，如分类广告、展品分类陈列、分类统计报表等。在网络环境下的数字化时代，类序法不仅广泛地应用于网络信息的排检，而且在计算机技术的支撑下，传统意义上的两种信息组织方法——分类法与主题法不仅相互渗透，且日益融合，实现了结构上的一体化。同时，搜索引擎的出现使得主题词排检法更加贴近普通用户。

6.2.1 学科体系分类排检法

由于人们一般都是在某个专业领域范围内从事科研、生产、教学、管理等活动的，习惯于从学科、专业角度出发来检索获取知识和信息。学科体系分类排检法是将词目、条目按其知识内容的学科属性，根据事先确定好的学科分类体系分门别类地加以归并集中，按一定逻辑顺序进行排检的方法。学科体系排检法不仅是信息管理领域处理各类信息资源的基本手段和方式，也是检索者按学科专业浏览和使用参考工具书的基本方法，很多百科全书、手册、年鉴等多采用此法编排条目。如有“工具书之王”称谓的百科全书，尤其是中文综合性百科全书（《中国大百科全书》），一般将信息条目先按学科分类编排，然后再按中文字顺（字序法）编排。像这种“粗线条”的学科体系分类排检法（指涉及学科名称而没层层细化学科内容）过于简单，此处不再赘述。而利用预先编制的分类法来分类排检信息是十分科学与复杂的。

^① 叶继元. 信息检索导论（第2版）. 北京：电子工业出版社，2009.

用此种排检法排列信息时,首先采用依照人为预先编制的分类法,采用其规定的学科专业概念及对应的系列分类号作信息内容标识,通过分类号的排序来排检信息。这种类序法将信息主题归入学科知识体系并使其号码化,借助所依据的分类法(分类体系)的自身结构(上下位类、同位类、交替类、相关类等)显示主题概念之间的派生、隶属和平行关系。学科体系分类排检法可以体现知识的学科属性和逻辑次序,可以实现学科信息的集中,便于按学科进行查找。

学科体系分类排检法通常选一种分类表(法)作为依据,也有不采用通行的分类体系而采用自编分类体系的,但都是按照学科体系归类的一种分类排检法。正如本书前面内容所述,古今中外,人们编制了形形色色的文献信息分类体系,包括图书分类法、档案分类法、专利分类法、标准分类法、资料分类法、公文分类法等。尤其是在20世纪以后,先后出现了以现代科学分类为基础的图书文献分类法,国外著名的有《美国国会图书馆分类法》(LCC)、《杜威十进分类法》(DDC)和《国际十进分类法》(UDC),它们是英语国家和地区使用最广的文献分类法。俄语类图书通常采用《ББК分类法》(《苏联图书馆——书目分类法》)进行分类。日语类图书通常按照《日本十进分类法》(にほんじっしんぶんるいほう, NDC)这一等级列举式分类法进行分类和排架。新中国成立后,陆续编制出版《中国人民大学图书馆图书分类法》(简称《人大法》)、《中国科学院图书馆图书分类法》(简称《科图法》)及影响最大和应用最广的《中国图书馆分类法》(简称《中图法》)。现当代这些学科体系分类法都是人工创制的用以表示文献内容信息及其相互关系的概念标识系统,它们主要依据概念的划分和概括的逻辑原理来建立自己的结构体系,以字母或数字或二者混合形成的代码生成分类号作为排检标识。

那么,学科体系分类排检法中的分类号(排检标识)是如何排序的呢?我们以《中图法》(第四版)为例,简单探讨一下图书文献被分类标引的分类号即排检标识的排序问题。例如,《语言美学》标引为H0-05;《孟子评传》标引为B222.55;《安徽当代书法集》标引为J292.88;《机织物的印花》标引为TS194.64;《清代草书》标引为J292.34=49;《列宁论教育》标引为A267;《全唐诗索引》标引为I222.742-7;《设计辞典》标引为TB47-61;《两宋财政史》标引为F812.944;《自然丛书》标引为N51;《激光探测大气污染》标引为X831;《文史哲》标引为C0;《标准化理论与实务》标引为G307.0;《大宅门》(随书附VCD一张)标引为I235.2;《全国报刊索引数据库》(光盘)标引为G216-794。对这些标引文献的分类号排序,换言之,这些分类标识在信息检索系统中的逻辑顺序应是:A267;B222.55;C0;F812.944;G216-794;G307.0;H0-05;I222.742-7;I235.2;J292.34=49;J292.88;N51;TB47-61;TS194.64;X831。从中可以看出,《中图法》分类号的顺序是按大类以英文字母的先后顺序排列,同一大类内,再按阿拉伯数字从小到大顺序排列,其中的辅助标记符号(包括用于区分同类书采用的一种著者号码表)等详细规则请参见第5章内容。不同的分类法标记制度不同,但分类号的排序都是依据其编号规则规定的逻辑顺序排列的。

在文献信息机构,分类法的类目体系和分类号一般是组织文献和图书排架的依据,同时也是编制分类目录的工具和进行信息资源导航的手段,尤其是近些年研制的电子版分类法,通过层层展开的方式对类目体系进行显示,更增强了其功能的发挥。值得关注的是,分类法功能的不同及使用的信息系统不同,这些都对分类法的结构设计及编码制度等提出了不同要求,此处不再赘述。

学科体系分类排检法的突出不足在于:在检索者通过采用分类法组织的信息系统检索小而专深的主题又不知分类号的情况下,检索时要逐层浏览,检索效率低下。检索者须熟悉该系统所运用的分类法,包括其所规定的类目体系涵盖的主题概念及概念划分标准、类目排列规则、编码规则和标识排列细则等很多技术性问题,这对于普通的检索者来说是困难的。

因此，鉴于学科体系分类排检法是按人为规定的一套号码进行排检的，其直观性较差，所以一般不会对在检索系统中独立使用，都会配备其他检索途径的字顺目录或索引。

需要补充的是，我们将档案分类法放在学科体系分类排检法所依据的分类法中，这是基于图书、情报与档案都属于大文献概念。而严格说来，档案分类法与图书分类法在类目体系划分标准上是有区别的。一般认为，档案是由人类在社会实践活动中形成的各种文件材料转化而来的，是当时活动的原始记录，而图书是事后在档案基础上经过锤炼、加工而编写出来的更能反映系统科学的结晶，是积累和传播知识的工具。因而档案与图书、情报等文献的根本属性不同，使得档案的分类不同于图书等其他文献的分类。档案信息分类存取系统中，档案是依据《中国档案分类法》（国家档案局主持制定，1985年10月全国文献工作标准化技术委员会第五分会对其进行了审议，1987年12月档案出版社出版试行本。1997年出版第二版。）进行分类编排的。《中国档案分类法》是用来标引与检索档案的，它有《中华人民共和国档案分类表》、《新民主主义档案分类表》、《民国档案分类表》及《清代档案分类表》。其设置19个基本大类：A 中国共产党党务，B 国家政务总类，C 政法，D 军事，E 外交，F 政协、民主党派、群众团体，G 文化、教育、卫生、体育，H 科学研究，J 计划、经济管理，K 财政、金融、保险、审计，L 商业、旅游业、服务业，M 农、林、牧、渔业，N 工业，P 交通，Q 邮电通信，R 城乡建设、建筑业，S 环境保护，T 海洋、气象、地震、测绘，U 标准、计量、专利。每一大类下，再分设若干属类，一般设4级类目，也有的设3或5级类目，形成等级分明、次第清楚的分类系统。从中可以看出其分类体系并不是学科体系层面的，与《中图法》不同。这是由于《中国档案分类法》以人类从事的社会实践活动为基础，以档案内容所反映的社会职能分工作为类目划分的主要标准。分类法是以反映档案内容特征的类目为基础，并用分类号进行标引和检索的一种分类检索语言工具。由于档案分类法一般与馆藏结构无关，因而综合内容的档案文件或案卷可以赋予多个分类号。因此，我们认为档案分类检索系统中的分类排检法不是严格意义上的学科体系分类排检法。

6.2.2 事物性质分类排检法

事物性质分类排检法是按同一类事物范畴的性质进行分类排检的方法，即把信息知识内容按事物属性分类，把相同范畴的事物汇聚在一起。这种方法中的各事物概念之间没有严格的系统性，一般按实际存在的事物性质划分类别。古代的书、政书和现代的年鉴、手册及某些辞书等多采用此法编排。例如，上海辞书出版社（1983年）出版的《同义词林》，就将词语概念按性质划分为人、物、时间与空间、抽象事物、特征、动作、心理活动等12大类^①。

这种方法是我国第一部词典《尔雅》开创的。《尔雅》分为19篇，后面16篇是分类词汇。如“释亲”就是把有关家族关系的词汇汇集在一起加以解释。后来一些解释词语的书也沿用这种排列方法。《尔雅》开创的体例也为后代编纂类书、政书所借鉴。类书按所采事、文的内容分门别类编排材料。其缺点是，由于对事物认识的局限性，某些事物的归类在今天看来并不恰当，因此查找起来会发生困难。例如，有关桥梁的材料，《艺文类聚》列入水部，《古今图书集成》列入考工典；有关薪、炭的材料，《艺文类聚》列入火部，《古今图书集成》列入草木典^②。古代的分类法对于今天来说，不仅存在着思想性和科学性上的问题，查检起来也颇觉不便。但为了使用它，必须仔细了解各级类目的含义，同时注意利用后人补编的辅助索引。

① 张帆，等．信息存储与检索．北京：高等教育出版社，2003：440．

② 冯惠玲，王立清．信息检索教程．北京：中国人民大学出版社，2004．

《中国古代名句辞典》等工具书同样采用此法编排。社会生活中许多信息的组织中, 有很多都是按所收物品和事物性质、用途来编排的, 如《机械零件手册》、《中国汽车车型手册》等。

由于所收集的信息所处领域不同、收集范围的不同、功用不同及古今分类标准的不一致, 资料的归类相差较大, 使用按分类法编排的工具时, 应先了解所用分类体系, 以确定所查资料的具体类目, 并注意相关类的查阅。

6.2.3 网络信息分类排检法

大多数网站、搜索引擎参照文献分类法的形式或类目体系, 按事物性质设计分类导航体系, 建立分类目录。例如 Yahoo! 的分类目录, 它按字母顺序将所有普通信息分成十四大类: 文化艺术、商业经济、计算机与 Internet、教育、娱乐、政府、健康、新闻媒体、体育、国家与地区、自然科学、社会科学、参考信息、社会文化。每个大类下面有许多子类及其精炼的描述, 每个子类又有数以千计的相关 Internet 网点信息。上述主页上的每一主题分类均是超链接词, 用鼠标点击任一个超链接词即可进入相应的页面。从整体上看, 目前的网络信息分类是一种综合性的信息分类。它不仅运用了现代文献分类的思想, 还吸收了事物分类的方法^①。其分类组织信息的做法主要是采用自动分类技术, 网页由机器人程序自动采集, 凡匹配预先规定类目的网页, 由计算机建立索引, 自动生成目录树。很多网站, 从不同角度对信息资源进行多重分类导航, 如一些政府网站, 将信息资源从机构、主题、服务对象、生命周期、体裁及信息格式等多种角度从事物性质类别上进行揭示与提供检索。显然, 普通网站网络信息分类面向应用的特点比文献分类要明显得多, 即其分类体系是将信息内容特征、网民需求特征及信息载体特征交叉混合构建而成的, 从科学角度看, 严密性不足, 实用性突出。因此, 不同的网站由于功能不同、信息收罗的范围和类型不同、面向的群体不同, 使得网络信息分类到目前还没有一部统一的分类法。

目前国内外很多关于将传统分类法应用于网络的研究都已进入实用阶段。有些网站采用了现有的文献分类法(有的经过一定的改造), 其应用范围主要是图书馆、主题网和各类学术性网站等。例如, 目前,《中图法》不但被图书馆用来类分图书和编制藏书目录, 而且被广泛应用在网络信息的组织中。例如, CNKI 就是利用《中图法》来提供分类检索的; 中国教育系统依照《中图法》对中国教育科研网上的信息资源进行分类; OCLC 根据 DDC 对 Net First 进行分类, 等等。学术性站点应用已有的文献分类法排检信息, 其分类目录多数没有分类号, 一般只是利用分类法体系建立浏览结构。其自动索引结构便于用户在查找时进行浏览, 提供检索主题的上下文。当用户检索目的不明确或检索词不确定时, 分类浏览方式更有效率。

Internet 上的更多站点信息分类组织是以事物为中心列类。其分类排检的主要特点有以下三点。

(1) 直接以语词组织信息, 未采用传统分类法的标记符号表达信息主题, 直接以语词表达类目体系, 采用链接技术链接网络文献, 比使用分类标记更加方便直观、易于理解。

(2) 排列方式简便。对同位类的排列, 主要有三种方式。① 字顺方式, 即, 同一上位类区分出来的类目按字顺方式排序; 每一类目下分出的子目再按下级类目的字顺排列, 形成一个层层展开的字顺分类系统。此种方式虽然未能揭示同位类之间的内容联系, 但是方便用户查找特定类目。② 以检索频率确定同位类的排列次序, 在同位类中首先列举高频类, 突出热门主题, 方便多数用户使用^②。③ 按照信息资源产生或入库时间排列。

① 储节旺, 郭春侠, 吴昌合. 信息组织学. 北京: 清华大学出版社; 北京交通大学出版社, 2007: 84.

② 周宁. 信息组织. 武汉: 武汉大学出版社, 2004: 69.

(3) 类目体系不稳定, 动态性强。由于网络信息的动态性, 使得网络分类体系在通过链接与网络信息建立联系时, 根据网络资源的发展变化会及时增设新类或重复反映相关类目。

在利用以事物为中心的网络分类排检法检索信息时, 有些方面需要注意: 如分类体系缺乏系统性和完整性、类目设置缺乏规律性、归类不科学、部分类名不确切、知识覆盖不全等。

6.2.4 主题词排检法

主题词排检法是以表征事物名称或概念的词语字顺为依据对信息进行编排、查检的方法。

结合第5章内容, 以及从网络信息分类排检法中也可以看出, 分类法与主题法虽然在组织信息资源时是两种不同的信息组织工具, 但原理上是共通的。它们在对同一文献进行揭示时表达的主题概念基本相同。分类法的检索标识(基本构件)是分类号, 主题法的检索标识(基本构件)是主题词, 而分类号与主题词在实质上都是代表一组相同主题文献(信息)的类集, 是文献(信息资源)所论述事物——主题概念的字面形式。可以说, 现代分类法在不同程度上都采用了按事物集中, 即以事物为纲, 按照研究对象排列的分类方法。在分类类目的设置上, 多是在前两、三级设置一些学科性的类目, 余下则主要面向事物及事物属性设置各种类目。从这个角度分析, 分类法与主题法类集的内容是相同的, 表达的都是主题概念, 处理的对象都是语义单元, 两者的差异只是在于类集方式不同和采用的标识不同, 基本构件的实质相同^①。例如, 《中国分类主题词表》就将分类表的类目名称(分类号)与主题词表中的主题词很好地对应起来, 它不仅是分类标引和主题标引的工具, 更是分类检索、主题检索及由分类号、主题词和自然语言三者组成的混合检索的有力工具。

主题词排检法是按表达信息内容的主题词来排检信息的, 其主题词是指对信息资源进行主题标引后的规范化的自然语词。作为标识符号的“规范化自然语词”——主题词, 是一种概括了信息资源的中心内容, 又用来标引和检索信息资源的标准词汇。正如学科体系分类排检法要借助于一种分类表一样, 主题词排检法中选词的参照体系是《主题词表》或机读库中的《主题词索引》。通常将以受控词(叙词、标题词、单元词)为存取标识所建立的信息存取系统称为规范词存取系统。规范词存取系统中的主题词排检法中, 主题词是排检的标识。其排列一般先依据主题词的字顺进行编排, 正如6.1节所述, 音序法、形序法等都可以成为编排主题词的方法; 然后再在主标题下排列副标题和次副标题; 同一主题词下的副标题的排列, 既可以采用字顺排列法, 也可以根据情况按标题之间特征关系排列; 相同标题下的款目可按信息资源的题名、责任者等其他特征名称排列。

主题词排检法一般不受学科领域层层划分概念的限制, 便于将不同学科专业、不同研究领域的相同主题信息集中一处, 提供按主题词字顺检索特定主题信息的途径。主题词排检法专指度高, 检索目标直观。主题词排检法在国外是比较常用的方法, 几乎每一种检索工具都有主题词排检途径。学术著作也大都附有主题词索引。在我国, 仅有部分书本式主题索引, 一些工具书的辅助索引采用了主题编排方法。主题排检法使用在社科中文工具书中并不多。这主要源于我国的文献检索长期以来是分类检索占主导地位, 而近二三十年来, 随着科学研究的交叉与深化、检索人员的非专业化、检索系统的计算机化, 主题检索的需求越来越多, 主题检索系统也就越来越普遍。尤其是随着互联网的蓬勃发展, 网络信息检索越发重要, 主题检索在我国有了很大发展。

主题语言在网络信息组织中的应用目前主要表现为关键词语言在网络搜索引擎中的广泛应用。国外大量标题词表和叙词表在图书馆网站的书目信息检索系统、网络联机数据库检

^① 戴维民. 信息组织. 北京: 高等教育出版社, 2004: 164.

索系统得到采用。目前还产生了一些网络化的主题词表,也称为联机词表(Online Thesaurus)。它可同时供网络用户使用,具有浏览和查询功能,查词比较方便。许多网络信息检索系统为建立特色服务,自行编制主题词表。例如,美国教育资源信息数据库使用的《ERIC 主题词表》、英国国家数字档案馆使用的《UNESCO 叙词表》和 UMI 数据库中的《ProQuest 受控主题词表》等。这些词表的作用主要是提供在线的实时帮助,不是用来信息标引的,而是检索的辅助系统,尤其是在一些专业性强、概念复杂、主题词拼写困难的数据库中,其辅助作用更为突出。目前,美国《国会图书馆标题表》(LCSH)和《医学标题表》(MeSH)已被一些网络信息检索系统采用。

6.2.5 网络信息关键词排检法

在互联网上,任何人都可以不受限制地自由出版、发布自己的网页,分布式存储成为网络环境中信息资源存在的主要形式,它区别于传统信息资源集中存储,又由于其信息海量和动态性强,难以有效控制。目前,对网络信息资源的组织管理有两种:① 依赖人工编制的主题目录,图书馆和信息专业人员通过对互联网的信息进行筛选、组织和评论,编制超文本的主题目录,这些目录虽然质量很高,但编制速度无法适应互联网信息的增长速度;② 依赖自动技术,计算机专业人员设计开发检索软件,对网页自动搜集、加工和标引。这种方式可向用户提供关键词、词组或自然语言的检索。因此,在网络与计算机时代,主题语言的优势得到了放大,表现出强大的生命力。其中关键词法在网络信息组织与检索中应用最为广泛。众所周知,关键词语言不是一种严格意义上的标引语言,但由于它在标引和检索中所发挥的类似于主题语言的作用,因而被视为一种准主题语言。在网络中,几乎每一个搜索引擎都具有关键词检索功能,这意味着搜索引擎的索引数据都采用了关键词法进行信息组织。

搜索引擎是一种由搜索软件、索引软件和查询软件共同完成搜索的网络信息资源查询系统(检索工具),一般由自动索引程序、索引数据库和检索服务三部分组成。它利用网络自动索引程序漫游网络信息空间,采集各站点信息,建立索引文档,形成数据库。数据库是其提供检索服务的基础,它搜集的信息以 Web 资源为主,还包含 Gopher、FTP、USENET、Mailing List 等网上信息资源。检索服务用来接收和解释用户的查询要求,然后根据一定的匹配策略,在索引数据库中进行查找,为用户显示按照某种相关程度排序的查询结果^①。通常看到的是其查询软件的用户界面。用户在用户界面的检索框中输入要搜索的关键词后,搜索引擎就通过自动搜索程序,从网站、网页的题名、地址、摘要,甚至网页的正文中抽取关键词作为索引词,提供指向相关网络资源的超文本链接。一般地,每一信息的关键词有多个,它们都是平等的,都按字顺轮流排至检索位置,都能作为检索入口。

不同的搜索引擎提供的关键词检索功能有所不同,有的只能进行简单关键词查询,有的既能提供简单关键词查询,又提供高级或复杂关键词查询。高级关键词查询包括:布尔查询、精确查询、模糊查询、截词查询、位置查询、字段查询、限制查询、管道查询、区分大小写查询和自然语言查询。

目前,由于网络的普及及获取信息的快捷与省时省力,搜索引擎成为人们获取网上信息的主要工具,其所提供的信息排序功能直接影响人们对信息的获取与利用。在 CNKI 初级检索说明里有输出排序说明(① 主题排序,即相关度;② 发表时间;③ 被引;④ 下载),其中的相关度就是相关性排序。相关性排序是指在检索到的结果集中能够先提供最有“价值”的网页给用户,这是体现搜索引擎优劣的一个重要指标。在实践中,人们采用多种手段

^① 徐天秀. 信息检索. 北京: 科学出版社, 2006: 150.

提高相关性排序的能力，包括：① 分析检索词在文中出现的位置，在标题和靠前的文字中出现往往具有较高的权值；② 为 META 置标中出现的词赋予较高的权值；③ 按照词频的统计规律加权；④ 按照 Web 站点排序^①。

搜索引擎对信息进行排序，目前主要运用 PageRankTM 技术（网页级别）、超文本匹配分析技术、内容相关度评价技术等。基于关键词的搜索引擎在决定网页的相关性排序时，一般遵循以下三大定律。

（1）地点和频率法。

地点和频率法最主要的算法就是看网页关键词出现的地点和频率。搜索引擎先检查标题中含有关键字的网页并认为它比其他网页的相关性更强。出现频率是搜索引擎决定相关性的另一个因素。搜索引擎会分析关键字在网页中出现的频率并与其他网页相比，关键字出现频率较高的网页被认为相关性更好。

（2）人气质量定律。

人气质量定律是搜索引擎的第二定律，它是受科学引文索引机制启发而提出的。科学引文索引机制认为被引用次数多的论文就是权威论文、好论文。那么在网上谁的网页被链接次数多，就认为该网页的质量高、人气旺。再结合相应的链接文字分析，就可以对检索结果排序了。Google、百度都采用了该定律。

（3）自信心定律。

人气质量定律解决的仅是技术层面的问题，然而搜索引擎融合了技术、文化和市场等各层面的因素。解决搜索引擎公司的生存和发展问题需要搜索引擎的第三定律——自信心定律。即向那些网站的拥有者们拍卖他们网站在检索结果中的排名，谁付的钱多，谁的网站就排在前面，且付费是根据网民点击该网站的情况来计算的，仅在检索结果中出现并不需要付费。根据这一定律，检索结果的相关性排序，除了以词频统计和超链分析为依据之外，更注重的是竞价拍卖。谁对自己的网站有信心，愿意为排名付钱，谁就排在前面。百度目前也采用这种排序方式。

6.3 时序法

时序排检法，简称时序法，又称年代排检法、编年排检法、纪年排检法。时序法是按照文献信息的写作、发表和出版年代或事物发生、发展的时间顺序或人物生卒年月日、生平经历的先后次序编排查检信息的方法。这种方法也是工具书编排方法中常用的一种。一些时间性较强的工具书，如年表、历表、大事记、史事纪年、记载个人生平的年谱、年鉴、个人著作目录等，都是按时序法编排的。

时序排检法按时间的顺序组合文献信息素材，即以信息的形成时间为排检标识，按时间顺序组织信息。大事记、传记资料、生平资料等多以时间为线索组织信息。这种方法能揭示信息内容的发展变化过程。按时间概念排列的工具书，便于人们查考历史时间，也便于来查考历史事物及图像资料。比较重要的检索工具书如《中国历史年表》、《中外大事年表》、《中华人民共和国经济大事记（1949年10月—1984年9月）》及《中国财政金融年表》等，均严格以时间先后为序编排资料，检索时只需按年索事，一查便得。

个人生卒年表、年谱及其著述目录，或采用顺时序法或采用逆时序法进行编排。时序法便于理清事物发展的脉络，从中可查考某些带有规律性的知识记录。

此外，时序排检法多用来作为社会科学文献检索语言的辅助方法，常用于辅助分类表，

① 都云程，卢献华．中文搜索引擎现状与展望．中文信息学报，13（3）：61～65．

起进一步细分的作用,如历史类按时序划分为古代史、近代史、现代史等。

要掌握时序法,尤其是在检索古代文献信息时,必须要熟悉古今中外纪年、纪月、纪日、纪时的常识及其相互间的换算。

6.3.1 历法常识

世界各国各民族在不同的历史时期记载时间的方法各不相同,各种文献中时间的记载又都以当时当地所用的历法为依据,因而在阅读和研究古今中外的文献时,必须注意时间的查检与换算。

世界各国的历法主要有三种:阳历、阴历,阴阳历。

1. 阳历

阳历,又叫太阳历,即现行历法,它曾经由罗马教皇八世格列高利于公元1582年修订,所以也叫“格列历”。“格列历”颁布第二年起陆续为世界大多数国家所采用,所以又称“公历”。又由于它始于欧洲,因而也称“西历”。我国1912年改用这种历法纪年,相对我国的原有历法,阳历又称“新历”。

根据天文历法,阳历是以回归年为基础的。地球绕太阳公转一周是365天5小时48分46秒(称回归年),即365.2422天,因此规定一年为365天,每隔三年加一个闰年(闰年2月29天)。将一年分为12个月,1月、3月、5月、7月、8月、10月、12月为大月,每月31天;4月、6月、9月、11月为小月,每月30天,2月是28天。阳历规定设置闰年,闰年为366天,而把含有365天的年份称为平年。置闰的规则可用三句话来表示:非世纪年的公元年数能被4整除的为闰年,世纪年(如1900年,2000年)的公元年数能被400整除的为闰年,其余的年份为平年。于是在400年内计有闰年97年,平年303年,平均长度为365.2425天,和回归年的长度只相差26秒,经过3000多年后才相差1天。这种历法与月亮圆缺没有关系,月数是人为规定的。

二十四节气是我国农历的一大特点。由于长期以来把农历称为阴历,因而不少人都误认为节气属于阴历,实际上节气完全取决于地球的公转。地球绕太阳运转一周约365天5小时多,运转94000万公里。这个公转轨道人们称为太阳黄经,分 360° ,我国古人把太阳黄经划成24等份,每份各占 15° ,为一个节气。两个节气间相隔日数为15天左右,全年即二十四个节气。规定太阳黄经等于零度(即 360°)时称为春分。从春分开始,依次为清明、谷雨、立夏、小满、芒种、夏至、小暑、大暑、立秋、处暑、白露、秋分、寒露、霜降、立冬、小雪、大雪、冬至、小寒、大寒、立春、雨水、惊蛰。因而有“太阳移至黄经 315° 时为立春”之说。由于太阳通过每等份所需的时间几乎相等,二十四节气的公历日期每年大致相同:上半年在6日、21日前后,下半年在8日、23日前后。并有两句口诀:上半年来六、二十一,下半年来八、二十三。但在农历中,节气的日期却不大好确定,再以立春为例,它最早可在上一年的农历12月15日,最晚可在正月15日。二十四节气反映了地球在轨道上运行时所到达的不同位置,可以说是阳历的一部分。

2. 阴历

阴历又称太阴历,是以月亮的圆缺、晦明的变化为基础来制定的。现在阿拉伯国家使用的回历就是阴历的一种。它规定1年为12个月,单月为大月,每月30天;双月为小月,每月29天,交替相间,以使历月平均长度接近于朔望月。由于月球绕地球一周需29.5306天,所以每30年置闰年11个,平年354天,包括6个大月和6个小月。闰年355天,在12月末增加一天,包含7个大月和5个小月。这种历法并未顾及到历年平均长度和回归年长度的配

合，久而久之，两者相差甚大。由于阴历的根本特点在于历月平均长度等于朔望月，每个日期就必然与一定的月相对应，如阴历十五大致就是满月而阳历的月是不能反映这一自然现象的。但阴历的历年则不能反映出季节的变化，和农业生产及人们的日常生活脱节，因而阴历已很少为人所用。

3. 阴阳历

阳历完全依据地球的绕日公转，阴历的历法完全根据月亮的运动，阴阳历则是两者并用。它同时考虑太阳和月亮的运动，把回归年和朔望月并列为制历的基本周期。由于阳历一个回归年是 365.2422 天，阴历的一年是 354.3671 天，这两种历法年每年相差 10 余天，因此每 19 年置 7 个闰年。凡闰年定为 13 个月。这种历法始于我国夏代，故称“夏历”，也称“中历”，俗称“阴历”或“农历”。

因而，我国沿用已久的农历并不是完全用阴历的，也不是完全用阳历的。一方面，农历以月亮绕地球运行一周为一月，平均历月长度等于朔望月；这一点与太阴历原则相同，所以也叫阴历；另一方面，农历设置闰月以使历年平均长度尽可能接近回归年，同时设置 24 节气以反映季节的变化特征。所以，农历是集阴、阳两历的特点于一身，所以有“阴阳历”之称。

阴阳历的历月长度和回历一样，有大小月之分：大月 30 天，小月 29 天。但农历历月的安排却不同于回历，回历中大小月机械地相间排列，而农历大小月则要经过推算后决定，比回历更加精密。农历规定月初必合朔，月朔之日定为初一。由于两个朔望月的长度并不正好为 59 天，因而一年中的大、小月数也不一定相等，有时可能连续出现两个大月或小月，以使历月的平均长度尽可能与朔望月相近，其剩余的差数则依靠闰月来调节。

朔望月和回归年是两个难以相合的周期，它们的余数都很零碎，而我国的农历却把作为阴、阳两历基础的这两个自然周期调和得十分成功。这也让我们领略到古人的聪明智慧。

6.3.2 中国古代的时序法

中华民族历史源远流长，随着封建政权的更替或科技的发展，纪年法就像古老的文明一样丰富多彩，主要有以下几种。

1. 王公在位纪年法

王公在位纪年法是我国最早的纪年法。这种纪年法按照一个国王或诸侯在位的年数纪年，依次称为元年、一年、二年、三年……直到离位时为止。殷商和西周时代以此纪年，如周平王四十九年、晋惠公元年。

2. 帝王年号纪年法

年号是帝王在位时用来纪年的名号。这种纪年法一般认为始于汉武帝建元（公元前 140 年）。汉武帝即位那年称为建元元年，其中“建元”就是年号。自此以后，历代皇帝都使用年号纪年。中间改换年号叫改元。明、清时代一个皇帝常常只用一个年号，所以人们用年号来称呼皇帝，如“崇祯皇帝”、“康熙皇帝”等。农民起义政权也使用年号纪年。这种纪年法一直用到清朝末年。史书记载中，采用年号纪年的同时还采用庙号、谥号、尊号等纪年。

3. 干支纪年法

天干和地支合称为干支。天干是甲、乙、丙、丁、戊、己、庚、辛、壬、癸的总称，是我国古代表示次序的符号，也叫十干；地支是子、丑、寅、卯、辰、巳、午、未、申、酉、戌、亥的总称，也是古人表示次序的符号，也叫十二支。十干和十二支循环相配，互相错综

组合纪年，可组成六十对干支，因而称作“六十干支”，因是错综组合，故名“六十花甲子”（通常也用来以指代六十岁，即花甲之年）。如表 6.2 所示。

表 6.2 六十花甲子

1 甲子	2 乙丑	3 丙寅	4 丁卯	5 戊辰	6 己巳	7 庚午	8 辛未	9 壬申	10 癸酉
11 甲戌	12 乙亥	13 丙子	14 丁丑	15 戊寅	16 己卯	17 庚辰	18 辛巳	19 壬午	20 癸未
21 甲申	22 乙酉	23 丙戌	24 丁亥	25 戊子	26 己丑	27 庚寅	28 辛卯	29 壬辰	30 癸巳
31 甲午	32 乙未	33 丙申	34 丁酉	35 戊戌	36 己亥	37 庚子	38 辛丑	39 壬寅	40 癸卯
41 甲辰	42 乙巳	43 丙午	44 丁未	45 戊申	46 己酉	47 庚戌	48 辛亥	49 壬子	50 癸丑
51 甲寅	52 乙卯	53 丙辰	54 丁巳	55 戊午	56 己未	57 庚申	58 辛酉	59 壬戌	60 癸亥

这个表通常称为“甲子表”。如甲子为第一年，乙丑为第二年，丙寅为第三年……六十年为一周。一周完了，再由甲子年起，周而复始，循环下去。我们在日历上看到的己巳年、庚午年，就是按干支纪年这种方法排列下来的。阳历年份除以 60 的余数减 3 便得该年农历干支序号数，再查表 6.2 的干支便是干支年纪。如果序号数小于、等于零，则在干支序号数上加 60。例如，求 1949 年干支， $1949 \div 60 = 32$ 余 29，年干支序号数 $= 29 - 3 = 26$ ，查干支表知该年为己丑年。1949—2009 年正好 60 年，那么 2009 年也是己丑年。同样可以算出 1978 是戊午年、2008 年是戊子年。

干支纪年萌芽于西汉，始行于王莽，通行于东汉后期。汉章帝元和二年（公元 85 年），朝廷下令在全国推行干支纪年。自此以后干支纪年和年号纪年并用。

4. 岁星纪年法和太岁纪年法

岁星纪年法和太岁纪年法是战国时期的以天象为基础的纪年法。岁星就是木星，在天体中运行一周约 12 年。用木星的这种运行规律来纪年就是岁星纪年。天文学家把木星的运行轨道分为 12 等份，叫 12 次，并且按顺序起 12 个名字：星纪、玄枵、娵訾、降娄、大梁、实沈、鹑首、鹑火、鹑尾、寿星、大火、析木。木星每年行经一次，称为一年。假如某年岁星运行到玄枵范围，这一年就记为“岁在玄枵”，第二年则记为“岁在娵訾”，其余依此类推，12 年周而复始。

由于岁星由西向东运行的方向与人们所熟悉的十二辰（“辰”本意指日、月的交汇点。“十二辰”则为夏历一年 12 个月的月朔时太阳所在的位置。《左传·昭公七年》：“日月之汇是谓辰。”命名上沿用地支，即子、丑、寅、卯、辰、巳、午、未、申、酉、戌、亥。）的方向和顺序正好相反，所以岁星纪年法在实际生活中应用起来并不方便。为此，古代天文占星家便设想出一个假岁星，叫太岁（又叫岁阴、太阴），让它和真岁星“背道而驰”，这样就和十二辰的方向顺序一致，并用它来纪年。这就是太岁纪年法。后来又起了 12 个太岁年名（又称岁阴名），岁阴的排列依次为：困敦、赤奋若、摄提格、单阏、执徐、大荒落、敦谿、协洽、虞滩、作噩、阏茂、大渊献。大概在西汉年间，历学家又起 10 个与天干相对的名称，叫岁阳，岁阳的排列依次为：阏逢、旃蒙、柔兆、疆圉、著雍、屠维、上章、重光、玄默、昭阳。然后岁阳、岁阴相配，同样可以排列出六十个“甲子”来。但因岁星运行一周只有 11.8622

年, 不足 12 年, 于是从东汉时起就不用岁星纪年法而改用干支纪年了。汉成帝末年, 由刘歆重新编订的三统历使太岁纪年和干支纪年从太始二年表面一样。虽然后来文献中仍有用它来纪年的, 那只是干支的别称罢了。如一民国年间书画作品印章边款上有“柔兆摄提格岁”字样, 则可推算出是“丙寅年”, 如果作者是民国的, 则即公元 1926 年^①。

5. 生肖纪年法

十二生肖也被称为十二年兽。在中国的历法上有十二只年兽依次轮流当值, 所以我们的中国年就有以鼠、牛、虎、兔、龙、蛇、马、羊、猴、鸡、狗和猪应用在历法上。即常说的子鼠、丑牛、寅虎、卯兔、辰龙、巳蛇、午马、未羊、申猴、酉鸡、戌狗、亥猪。生肖纪年法就是用 12 种动物名称 (叫十二生肖或十二属相) 与 12 地支相配来纪年的一种方法。12 年一循环。其相配情况如表 6.3 所示。

表 6.3 生肖纪年法

地 支	子	丑	寅	卯	辰	巳	午	未	申	酉	戌	亥
动 物 名	鼠	牛	虎	兔	龙	蛇	马	羊	猴	鸡	狗	猪

这种纪年法早在东汉王充的《论衡》、许慎的《说文解字》中就有记载。

6. 古代纪月法

我国古代纪月法主要有三种。其一是数序纪月法, 是我国最早的纪月法。即 1~12 月的月数, 岁首称正月, 秦时又曾改称为端月。第二种是地支纪月法 (十二地支)。古人常以十二地支配称十二个月, 每个地支前要加上特定的“建”字。如杜甫《草堂即事》诗: “荒村建子月, 独树老夫家”, “建子月”按周朝纪月法指农历十一月。第三种是时节纪月法。如《古诗十九首》: “孟冬寒气至, 北风何惨栗。”“孟冬”代农历十月; 陶渊明《拟古诗九首》中“仲春遘时雨”, “仲春”代农历二月。此外, 还有花木纪月法 (如一月杨、二月杏、三月桃、四月槐、六月荷等)。

7. 古代纪日法

我国古代纪日法主要有以下 4 种^②。

1) 数序纪日法

《梅花岭记》: “二十五日, 城陷, 忠烈拔刀自裁。”归有光《项脊轩志》: “三五之夜, 明月半墙。”其中的“三五”指农历十五日。

2) 干支纪日法

干支纪日法就是每天用一对干支表示, 逐日记录, 60 日后重复。这种纪日法是我国古代历法中很重要的组成部分, 而且使用非常早, 可以说是我国最古的一种纪日方法。如《诗经》、《春秋》、《二十四史》、《清史稿》、《资治通鉴》等皆用于支纪日。据初步统计, 最早是从春秋鲁隐公三年 (公元前 720 年) 二月己巳日起连续纪日, 一直到清代宣统三年 (公元 1911 年) 止, 计有 2600 多年的历史, 这是迄今所知世界上最长久的纪日资料。如《左传·僖公三十三年》: “夏四月辛巳, 败秦军于殽。”其中的“四月辛巳”指农历四月十三日。苏轼《石钟山记》: “元丰七年六月丁丑。”“丁丑”即农历六月九日。姚鼐《登泰山记》: “是月丁未。”“丁未”指这个月的十八日。古代还单用天干或地支来表示特定的日子。《礼记·檀弓》: “子卯不乐。”其中的“子卯”代指恶日或忌日。

① http://www.news365.com.cn/wxpd/jc/hsshmj/zxt/200702/t20070202_1279815.htm (2009-8-14)。

② <http://www.yuwen99.cn/gaokao/ShowArticle.asp?ArticleID=42519> (2009-8-14)。

3) 月相纪日法

月相纪日法是指用“朔、朏、望、既望、晦”等表示月相的特称来纪日。每月第一天叫朔，每月初三叫朏，月中叫望（小月十五日、大月十六日），望后这一天叫既望，每月最后一天叫晦。例如苏轼《赤壁赋》：“壬戌之秋，七月既望。”

4) 干支月相兼用法

干支置前，月相列后。例如姚鼐《登泰山记》：“戊申晦，五鼓，与子颖坐日观亭。”

8. 古代纪时法

我国古代主要有如下两种计时法^①。

(1) 天色纪时法。

古人最初是根据天色的变化将一昼夜划分为十二个时辰，它们的名称是：夜半、鸡鸣、平旦、日出、食时、隅(yú)中、日中、日昃(zè)、晡(bū)时、日入、黄昏、人定。

(2) 地支纪时法。

以十二地支来表示一昼夜十二时辰的变化。近代又把每个时辰细分为初正，这就等于把一昼夜分为24小时。由于午时是现在的11:00~13:00，所以人们就称“午时”中间的12:00为中午或正午，称天亮到12:00以前为上午，12:00以后到黄昏以前为下午。

古天色纪时、地支纪时与现代的钟点序数纪时对应关系表如表6.4所示。

表6.4 古天色纪时、地支纪时与现代的钟点序数纪时对应关系表

天色	夜半	鸡鸣	平旦	日出	食时	隅中	日中	日昃	晡时	日入	黄昏	人定
地支	子	丑	寅	卯	辰	巳	午	未	申	酉	戌	亥
现代钟点	23:00	1:00	3:00	5:00	7:00	9:00	11:00	13:00	15:00	17:00	19:00	21:00
	~ 1:00	~ 3:00	~ 5:00	~ 7:00	~ 9:00	~ 11:00	~ 13:00	~ 15:00	~ 17:00	~ 19:00	~ 21:00	~ 23:00

我国古代把夜晚分成五个时段，古时常夜间击鼓报更（古时用滴漏计时，夜间凭漏刻传更），所以古人常以鼓代更，所以叫做五更、五鼓，或称五夜。每更分为五点，每点约等于现代的24分钟。如《孔雀东南飞》：“仰头相向鸣，夜夜达五更。”《群英会蒋干中计》：“伏枕听时，军中鼓打二更。”《李愬雪夜入蔡州》：“四鼓，怨至城下，无一人知者。”夜间钟点与时辰对应表如表6.5所示。

表6.5 夜间钟点与时辰对应表

夜间时辰	五更	五鼓	五夜	现代时间
黄昏	一更	一鼓	甲夜	19:00~21:00
人定	二更	二鼓	乙夜	21:00~23:00
夜半	三更	三鼓	丙夜	23:00~1:00
鸡鸣	四更	四鼓	丁夜	1:00~3:00
平旦	五更	五鼓	戊夜	3:00~5:00

更尽天明，进入卯时（早晨5:00~7:00）。古人上朝或去官府当差就在这个时辰。官家查点到的人数，这叫点卯。现在有些人戏称上班应付差事为点卯或应卯，就是沿用这个典故。“三更”“五更”的“更”读gēng。过去北方方言中有读jīng的，词典也曾保留过这个旧读音，

^① <http://www.saybest.cn/baike/zhexue/21701.htm> (2009-8-14)。

现在已经废除了。此外，古人把一昼夜分为100刻，实算96刻，每刻15分钟。漏刻指很短的时间。

6.3.3 时序法的应用

利用古代按时序法编排的文献及阅读古籍，首先要熟悉古代的计时方法。目前世界上通行的公元纪年中，把传说中的基督教创始人耶稣诞生的那一年作为公元元年。这一年相当于我国汉平帝元始元年。这以前的年份称“公元前”，这以后的年份称“公元”。在阅读、研究，特别是注释古文献时，常常需要把上述的古代纪年转换成公元纪年。

时序法常用于一些时间概念比较强的参考工具。如年表、大事记、年谱、人物传记等。利用按时序法编排的工具书进行检索时，如利用“生卒年表”或“年谱”来查考人物资料时，需要辅以人名索引才能使用。例如，利用《历代人物年里碑传综表》，即先查人名字顺索引后再查所需的人物事迹。

此外，时序法在强调时间性的信息排检中尤显重要。例如，在编制和使用地方文献索引数据库中，时间排序法就有重要作用。在输入的每条索引数据中都包含一个时间项，以便尽可能准确地标出地方文献中事件发生的时间。例如，历史人物和革命领导干部要标引其到达某地或在某地任职的时间；劳动模范、先进人物要标引出被授予荣誉的时间；优秀运动员要标引其创造纪录或夺取名次的时间；科学家要标引其研究成果被评定的时间；革命英雄要标引其在某地参加革命的时间；文章、讲话及著作要标引其出版、发表的时间；等等。在索引数据库中时间成为一个检索点，并具有排序功能。它可使数据库中的数据按时间次序排列，形成一部大事记的资料线索。例如抗日战争史料，按时间排序后对研究抗日战争历史有很大益处。社会生活中以时序法排检信息的例子比比皆是，如手机接收信息、电子邮箱接收来信都是自动按照时间倒排顺序呈现的，一般网站新闻网页、Blog日志等，都是以时间倒排方式来排列的，最新信息都在最上面。按时间倒排顺序的方法是与信息的时效性紧密联系在一起。

总而言之，现代的时序法编排方法比较简单，用户可根据时间顺序（顺查或逆查）检索所有信息；用户在使用古代时序法时应注意各纪年/月/日法的换算。

6.4 地序法及其他排检法

任何事情都是在特定的时间与空间中产生、存在、运动着的，正如时序法一样，空间特征同样可用于排检信息的依据。地序法也是信息排检法中常见的一种。同时我们也看到，社会生活是丰富多彩的，社会实践中因信息的类型、时效、功用、需求情况等不同而产生出各种各样的信息排检方法。但是无论采取怎样的排检法，目的都是科学编排，方便用户检索、获取和利用信息。而随着计算机、网络技术的发展，排检法在软件程序的支撑下也在悄然地发生变化，用户可以用自然语言轻松检索，检索结果在一定程度上可以按用户喜好排序。

6.4.1 地序法

地序法即地序排检法，是指按照信息中所涉及地理位置或行政区域名称为标识来排检信息的方法。它主要用于地域特征比较明显的信息系统或工具书中。这种方法可以把同一地域的有关信息素材集中在一起，全面地反映某一地区、某一国家、城市、乡镇等的历史和现状，如《中国地方志综录》。

地序法以信息的形成地区或信息内容所反映的地区为序化符号，按行政区划排列法来组

织信息。利用地序法进行信息排序时,一般有序可循。如果是国际性的,或者先区分洲,而后依地理位置从北到南、从西向东排列;或者按国家名称的字母顺序排列。如果是某一国家的,通常以该国规定的行政区划为序。这种方法能反映有隶属关系和横向联系地区的信息。

地序法主要用于编制和检索地理和地方资料的工具书。它可以用在研究查考自然资源及经济开发的工具书中,如编制地图集、有关地理资料、方志目录、地方资料等工具书;各类图书中凡涉及世界各国和国内各地区的,也都采用地序法。例如《中国名胜词典》、《中华人民共和国分省地图集》、《历代地理沿革表》、《中国边疆图籍录》、《合肥市地方志》、《中国邮政编码图集》和《欧洲金融年鉴》及《中图法》等分类法中的《地区复分表》均按地序法编排有关资料。这些工具书多数附有地名索引,以便在不知地名所属地域时按地名顺序查找。在有些情况下,地序法与时序法交替使用以形成多层次的信息集合,如地方志等。

6.4.2 其他排检法

信息排检法的种类如同信息内容的多样性一样,丰富而无穷。本章前述的信息排检方法都是在信息组织与信息检索领域中最常见的,在此简单介绍几种其他排检法。

1. 谱序法

谱序法是按照机构建制、血缘关系依次编排文献的方法。常见的检索工具如《历代职官表》(清,纪昀等编,上海古籍出版社1989年影印本)。其所列的76个表即以清代官制为纲,从中央到地方逐级排列各政权机构的职官,附官名索引,是按官名查检的工具书;世袭表和族谱则按照血缘关系依次排列,如洪秀全家的《洪氏宗谱校补本》(1981年版)^①。

2. 代码排序法

代码排序法是在某一社会领域采用有一定行业或专业含义的代码来序化信息的方法。如用邮政编码组织信件,用身份证、学号组织人群信息,用地址码(门牌号)组织住户信息等。文献编号(如专利号、报告号、标准号等)在组织各类信息方面有着重要作用。这类序化法在特定的专业领域是专业信息序化的重要方法。代码排序法具有简单易用、唯一、标准等特点。

3. 引证关系排序法

引证关系排序法利用信息之间的引证与被引证关系来组织信息。如利用文献之间的引证关系所组织的各种引文索引系统。引证关系排序法不仅是组织信息的一种独特方法,而且是进行各种评价研究的基础。

4. 权值排序法

权值排序法即赋予不同信息以不同的权值,以权值大小为依据组织信息的方法。实质就是按照信息的重要性大小来组织信息的方法。如决策方案的选择、教学质量的评估等都涉及权值组织法。甚至报纸在版面安排上,最重要的信息总是放在头版头条的位置。再如电视节目目的安排,总是把重要节目放在黄金时间播出。

5. 其他标准排序法

根据某类用户、某一用户或用户某一方面的特殊需求组织信息的方法,如股票信息、旅游信息、证券信息等。利用信息中其他特性如颜色、重量、速度等其他特征来组织信息,等等。

众所周知,信息存储系统的信息要满足不同使用者的需求,因此,要充分考虑不同使用者的能力和需求,尽量运用多种排检方法以有利于满足不同查阅者的需求。《新华字典》所

^① <http://xxjs.col.ynu.edu.cn/dzjy/1.doc> (2009-8-14)。

收的字是按音序法排的，但还列有《部首检字表》（包括《部首目录》、《检字表》和《难检字笔画索引》），这样做是比较合适的。

6.4.3 计算机程序与动态信息排检法

计算机技术、网络技术与数字化技术的发展给信息组织技术带来了变革，信息组织技术的变革直接影响着信息的检索与获取，超文本链接技术使人们可以在网络中任意遨游。各种计算机软件让编排、检索信息变得轻而易举。

1. 超文本链接

互联网是目前最大的信息资源网，它之所以能够使这些信息资源为广大用户所利用，主要依靠下面的三项基本技术：① 确定网上信息资源标识的统一命名方法，URI（Uniform Resource Identify）包括统一资源地址（URL，Uniform Resource Locator）和相应的路径与文件名；② 存取资源的网络协议：超文本传送协议（HTTP，HyperText Transfer Protocol）；③ 在资源之间很容易跳转、浏览的超文本链接技术（HyperLink）^①。正是这三项技术，使人们可以通过输入网址直接浏览站点信息；通过搜索引擎检索网络信息。

超文本（HyperText）的基本结构由节点（Node）和链（Link）组成。节点用于存储各种信息，链用于表示各节点（即各知识单元）之间的关联。通常的文本信息是用字符串来表达并以线性方式顺序排列的，这种编排方式并不完全符合人们的思维习惯，因为人类的思维很少是线性的，更多是联想式、跳跃式的，是在多角度、多层次上同时展开的过程。而超文本是以非线性方式组织的^②。这里的“非线性”是指文本中遇到的一些相关内容通过链接组织在一起，用户可以很方便地浏览这些相关内容。这种文本的组织方式与人们的思维方式和工作方式比较接近。

超文本链接（HyperLink）是指文本中的词、短语、符号、图像、声音剪辑或影视剪辑之间的链接，或者与其他的文件、超文本文件之间的链接，也称为“热链接（HotLink）”，或者称为“超文本链接（HyperTextLink）”。词、短语、符号、图像、声音剪辑、影视剪辑和其他文件通常被称为对象或称为文档元素（Element），因此超链接是对象之间或文档元素之间的链接。建立互相链接的这些对象不受空间位置的限制，它们可以在同一个文件内也可以在不同的文件之间，也可以通过网络与世界上的任何一台联网计算机上的文件建立链接关系。

因此，超文本存取系统是一种网络式的“数据库”，它是按照人脑的联想思维非线性存储、管理和浏览信息的网状结构。在超文本检索系统中，当信息资源数量庞大时，链接的线路会十分曲折，用户容易迷失方向。

2. Tag 技术与信息排序

Tag，直译为标签，是一种由用户自定义的、用于描述信息的关键词，特点是无层次结构，自定义，利于普通搜索查找。Delicious、Flickr 等 Web2.0 网站的发展促进了它的流行，使之成为社会化书签、相册服务、博客等网站的常见功能。但是 Tag 也不同于一般的关键词，用关键词进行搜索时，只能搜索到文中提到的关键词，但 Tag 却可以将文中根本没有的关键词作为 Tag 来标记，并可以将所有带有同样标签的文章全部关联起来，便于查找。从目前网络发展形势而言——个人自生成内容（博文、图片、视频等）迅速增加，各类信息海量涌来——Tag 可谓是信息管理的强大利器。

① 周宁. 信息组织. 武汉: 武汉大学出版社, 2004: 180.

② 冷伏海. 信息组织概论. 北京: 科学出版社, 2003: 19.

在启用了标签技术的信息系统中,信息发布者可以在附设的栏目里为其所发布的信息(文本或图片等)添加一个或多个代表其内容的词语——标签(Tag)。在启用了标签技术的交互性网站中,每一个用户都可以是标签的创建者,系统会通过相应机制将标注有相同标签的信息聚合在一起,将达到一定词频的标签排列在页面上,一般按字典顺序排序,用字体大小和颜色来表示标签的热度。标签具有超链接功能,可以把信息检索者引导到标有同样标签的信息的导航页面上,这样检索者就可以快捷地找到一系列与其感兴趣主题相关的信息,而且,标签还建立了与信息发布者之间的关联,可以帮助有共同兴趣和关注点的不同检索者建立联系和交流,形成网络社区。^{①②}

3. 应用软件与信息排序

各种应用软件的信息排序功能也关系着人们对信息的编排与使用的效率,在此只略举一二。

1) 中文文件名按笔画排序

在 Windows 的资源管理器中,不管文件(或文件夹)名是英文还是中文的,当使用“详细信息”方式查看时,在文件列表标题“名称”上单击,文件名默认的排序方式都是按字母的顺序排列的。如果有大量的中文文件名,那么让中文文件名按笔画排序会更符合使用习惯。具体方法如下所述。

在 Windows XP 系统中,双击“控制面板→区域和语言选项”,切换到“区域选项”选项卡,单击“自定义”按钮,弹出“自定义区域选项”对话框,打开“排序”选项卡,在排序方法下拉列表中选择“笔画”。重新启动计算机后,打开资源管理器,单击文件列表标题“名称”,会发现中文文件名已经按笔画多少排序了!

注意:该设置只影响中文名称文件,不管以“发音”还是以“笔画”排序,用英文命名的文件,其排序方式总是按名称排序的。

2) Excel 让数据按需排序

如果你要将员工按其所在的部门进行排序,这些部门名称的有关信息既不是按拼音顺序的,也不是按笔画顺序的,怎么办?可采用自定义序列来排序。

(1) 选择“格式→选项”命令,弹出“选项”对话框,进入“自定义序列”选项卡中,在“输入序列”下面的文本框中输入部门排序的序列(如“机关、车队、一车间、二车间、三车间”等),单击“添加”和“确定”按钮退出。

(2) 选中“部门”列中任意一个单元格,选择“数据→排序”命令,弹出“排序”对话框,单击“选项”按钮,弹出“排序选项”对话框,按其中的下拉按钮,选中刚才自定义的序列,按两次“确定”按钮返回,所有数据就按要求进行排序了。



本章小结

排检法应用十分广泛,如检索工具的编制和检索系统的建立,工具书的编制,文献的排架,电话簿、各种名录的编制等;反过来,我们对这些工具书、数据库等信息系统进行检索等也都需要掌握信息排检法。常用的排检法有分类排检法、字顺排检法、号码排检法、时序排检法、地序排检法等。信息排检法种类繁多,各有千秋。在信息组织时,字序法是编排字

① 于明洁. 豆瓣网 Tag 模式对图书馆信息组织的启示[J]. 数字图书馆论坛, 2009(12): 99~102.

② 邓卫华等. 虚拟社区中基于 Tag 的知识协同机制——基于豆瓣网社区的案例研究[J]. 管理学报, 2012(8): 1204~1210.

词典的常用方法，而音序法具有国际通用性，部首法、笔画笔形法、四角号码等方法是我国汉语文字编排所特有的传统方法。分类法、主题法、时序法和地序法等是从信息内容的特性方面编排信息的方法，它们在信息排检时并不是独立发挥作用的，还需要借助字序法。在计算机网络时代，主题法更有发展空间。在计算机程序的驱动下，数字化信息排序变得越来越随意、方便。



问题讨论

1. 信息排检法有哪些种类？
2. 部首法中如何确定部首？
3. 四角号码法中的笔形取码规则是怎样的？
4. 我国农历是哪种历法？二十四节气属于阴历吗？
5. 十天干、十二地支、十二生肖各是什么？
6. 主题法与分类法的区别与联系是什么？
7. 检索自己名字的部首、带附号四角号码和汉语拼音。
8. 检索自己生日的阳历、农历、干支历对照，并注明文献来源。
9. 清人袁枚在《祭妹文》的开头就写道：“乾隆丁亥冬，葬三妹素文于上元之羊山……”这里的“乾隆丁亥”是公元哪一年呢？怎样把古文献中的历史纪年改换成公元纪年？



第7章


信息组织成果与工具

内容提要

信息经过前期的选择、揭示、标引、描述、排序等环节，即可形成信息组织成果。其中，大量的信息组织成果又成为人们进行信息检索的工具。越来越多的检索工具的出现，使其又进入了信息组织领域。

本章以实例介绍那些使用相对广泛的、综合性的信息组织工具：目录、索引和文摘。同时，本章还论述了作为当前信息组织最主要成果的重要组成部分的全文数据库、数据库整合与导航以及搜索引擎。

本章重点

- 目录的类型；
 - 索引的概念和类型；
 - 文摘；
 - 全文数据库的开发步骤与技术；
 - 异构数据库整合；
 - 信息导航系统的建设；
 - 搜索引擎的类型和工作机制。
- 

7.1 目录

目录又称书目。英文 Bibliography 一词是由希腊文“Biblion”（book，书）和“Graphein”（Writing，写）两个单词融合而成的，其最初的含义是“图书的抄写”（The Writing of Books）。后来随着图书的增加，社会对图书概况记录需求的提高，该词的词义也逐步演变成“关于图书的描述”（Writing about Books）。现代意义上的目录的含义就是著录一批相关的文献，按一定的次序编排而成的一种登记、报道和宣传书刊文献等实体和虚拟信息的检索工具。

目录的名称繁多。中国古代多称目录，西汉刘向编撰的《别录》一书中，著录有《列子目录》，是“目录”一词的最早出处。还有其他名称，如“略”、“簿”、“录”、“书录”、“书录解題”、“题记”、“题识”，还有“考”、“经籍考”及“书目”、“总目提要”、“综录”、“总录”等。英文中也有用 Index、Guide、Record、Catalog 等词来表示目录这一工具的，如 Cumulative Book Index、Guide to Reference Books、American Book Publishing Record、National Union Catalog 等。

目录能反映一定历史时期科学文化发展的概貌，是人们对浩如烟海的文献进行记录、整理，并加以控制与管理，便于存储与传递的有效手段，也是查阅和利用文献必不可少的工具。

7.1.1 目录的类型

目录的种类很多，不同的划分标准下有不同的目录类型。国内外的划分方法也不统一。西方国家一般把书目分为三类：一种是列举式书目（Enumerative Bibliography），对书目信息做简要的描述；一种是描述性书目（Descriptive Bibliography），对文献的特征（著者、题名、出版项、页码、书型、插图等）做详细的描述；还有一种被称为评论性书目（Evaluative Bibliography），是对著者、成书年代及版本插图进行考证或物质特征组织资料。这些特征可供书史、版本研究的考证，相当于我国的版本目录。中国学术界根据编撰方式和时间等方面的特点，一般将书目分成古典书目和现代书目两种。古典书目包括官修书目（如《四库全书总目》）、史志目录（如《汉书艺文志》）、私撰书目（如《郡斋读书志》）及版本目录（如《遂初堂书目》）等。现代书目，按不同的角度可划分更多的类型，其中最常用的几种分类方式如下。

1. 按编制目的和社会职能分

（1）国家书目（National Bibliographies），是全面揭示和报道一个国家出版的所有文献的系统目录。其主要特征包括：所收录的材料主要是在这个国家内出版、印刷、发行的；用这个国家通用的语言编纂的；收录文献所揭示的内容大部分是同这个国家相关的；大部分内容是不断更新的。UNESCO 就国家书目的收录范围所提的建议是，国家书目应不仅收录所有出版社出版的书籍，而且应收录非书商经销的书籍；从形式上看，应包括图书、连续出版物、视听资料、地图、论文、艺术印刷品、电影胶卷、缩微资料等，以及如今的电子版文档等。国家书目编制工作一般由收藏丰富的国家图书馆承担，并由呈缴本制度保证，收录齐全、著录规范。在全世界 200 多个国家和地区，已有 90 多个国家拥有国家书目，较著名的有：《英国国家书目》、《法国和世界法语出版物总书目》等。美国没有呈缴本法案，也没有正式的国家书目，而代之以《全国联合目录》（NUC，National Union Catalog），由 1790 年的著作权登记制度作为支持，其收藏包括国会图书馆在内的北美 1100 个图书馆的藏书，堪称世界书目。国家书目包括回溯性国家书目和现行国家书目。回溯性国家书目是全面反映一国在一定历史时期内的图书文献的总目，如《民国时期总书目》；现行国家书目则全面报道一国近期出版的图书文献，如《中国国家书目》。

(2) 营业性书目 (Book Trade Bibliographies) 是一种为满足书业贸易需要、由出版商以赢利为目的而编纂的书目, 早年一般涉及一国范围的出版物; 而在目前, 一些大型的网络版书目数据库多已集中了多个国家的出版物。一些编纂质量好的营业书目在某种程度上可充作国家书目。营业书目主要用于商业目的, 因而出版信息齐全, 但不一定会有完整的编目信息。在形式上有征订目录、预告书目 (预报数周或数月内将要出版的书刊)、在版书目 (Books in Print) 等。在版书目亦称库存书目或在销书目, 反映在特定时间内市场上所供应的图书, 内容广泛, 出版及时, 是国家书目的补充。如鲍克公司出版的《美国在版书目》(BIP) 和《美国出版商目录年报》(PTLA)、惠特克公司的《英国在版书目》(BBIP)。

(3) 推荐书目, 亦称导读书目 (Best Books), 是针对特定的读者, 围绕某一主题, 选择、推荐有关文献, 用以指导阅读而编制的书目, 如《书目答问》。

2. 按收录内容及范围分

(1) 综合性书目 (Comprehensive Bibliographies)。它是将各个学科门类的图书汇总编成的一种图书目录。其内容广博、包罗万象, 既有人文社会科学方面的书, 又包括自然科学和应用技术方面的书; 层次也不同, 既有普及性读物, 又有学术性著作。国家书目和大部分营业书目属于综合性书目。

(2) 专题书目 (Subject Bibliographies) 是为某学科专业或某一研究课题编制的书目。它集中揭示某一方面的全部图书信息, 具有较大的使用价值, 从事科学研究应尽量使用专题书目。也包括那些专门收录某一作者的全部著述并兼收研究该作者的资料的个人著述书目, 以及那些专门收录有关某一地区历史、自然和社会状况的图书文献的地方文献书目。如《安徽文献书目》。

3. 按著录文献类型分

书目可分为图书目录、丛书目录、报纸目录、期刊目录、方志目录、乐谱目录等。

4. 按文献流通过程和环节分

书目又可分为出版发行目录和收藏目录, 其中, 收藏目录又可分为私人藏书目录和图书馆馆藏目录。而图书馆馆藏目录则又有单馆馆藏目录和多馆联合馆藏目录之分。联合目录是反映多个图书馆文献收藏的目录, 如《中国地方志联合目录》、《全国中文期刊联合目录》及美国《全国联合目录》(NUC) 等。这是信息资源共享的重要前提之一, 网络的普及为联合目录提供了优越的条件。

5. 按载体形式分

按载体形式不同, 可分为书本式目录、卡片式目录、计算机输出缩微目录 (COM, Computer Output Microfilm) 和联机公共检索目录 (OPAC, Online Public Access Catalog)。在互联网如此普及的今天, 人们对图书、期刊等文献的管理多采用计算机操作, 而很少用卡片目录了。书本式目录的用户也在减少。相反的是, 联机公共检索目录却日益普及。全世界任何联网计算机都可以查询馆藏目录, 从而知道哪个图书馆都收藏有什么图书或期刊, 还可以进行网上续借、预约、查询个人借阅情况等。

7.1.2 目录举要

1. 中文目录举要

我国有着悠久的编撰与出版目录的历史, 千余年来, 我们有着一系列反映从古至今文献

出版与收藏信息的目录。

1) 检索中文古籍的目录

古代图书通常称为古籍,在国内一般指辛亥革命以前的人所撰写的著作,以及后人整理的古代的文献,如影印本、校点本、汇编本等。我国现存古籍约有 10 万多种。要实现对这些古籍文献信息的检索与开发,就需要利用特定的书目。

(1) 查古籍流传。

查找古籍的流传情况,主要利用史志目录和公私藏书目录。东汉班固根据刘歆的《七略》编写《汉书·艺文志》,开创了史志目录的先河。此后史家修史,都仿效班固作艺文志或经籍志,如《隋书·经籍志》、《明史·艺文志》等。但二十五史中,并不是每部史书都有艺文志或经籍志。清代以后,一些学者相继对正史中的艺文志进行了补撰工作,包括补写、增补或注释等。由于艺文志或经籍志都是根据当时的政府藏书并参考其他官私书目编成的,因此,利用它可以了解古籍的流传情况。要了解某书在历代艺文志或经籍志中是否著录,目前比较常用的工具书是《艺文志二十种综合引得》(哈佛—燕京学社引得编纂处 1933 年编,中华书局 1960 年重印)。

要查考我国历代文献流传使用、评价的信息,一般可以利用《通志》、《文献通考》、《四库全书总目》和《书目答问》等传统目录。其中使用最多的是《四库全书总目》。由于该书篇幅浩大,不便查检,后又被删节成 20 卷的《四库全书简明目录》,仅收录了《四库全书》中所收的 3000 多种书,并浓缩提要。通过这一繁一简两套目录,大多数在我国清代中期以前存世的古籍可以被检索到。另外,《四库全书总目》中也存在一些错误、疏漏,因此要注意利用后人纠谬、补正之作。

(2) 查古代丛书。

因为战乱、灾害、年代久远或其他原因使得一些古籍失传,但有些古籍则因其被收录于丛书中而得以存世。从这个角度看,丛书对保存古籍具有重要的作用。查找古籍丛书最常用的工具书有:《中国丛书综录》、《中国丛书广录》、《中国丛书综录补正》,以及《中国丛书目录及子目索引汇编》等。台湾省的台北德浩书局 1974 年版的《丛书总目续编》对《中国丛书综录》第一册“总目”进行了补充,收录了我国台湾地区出版的丛书 600 余种,可查知我国台湾地区 1974 年前丛书的编纂、收藏情况。还有《四部丛刊书录》、《四部备要书目提要》、《丛书集成初编目录》等,都可作为查找古代丛书之用。

(3) 查古籍版本。

版本是一书在流传过程中形成的各种本子。如书写或印刷的形式、年代、版次、字体、装订、内容的更改等,都形成不同的版本。常用的版本目录有:《增订四库简明目录标注》、《中国古籍善本书目》、《中国善本书提要》等,还可以利用《书目答问》、《贩书偶记》、《中国丛书综录》等。

2) 检索近现代中文图书的目录

若要查找我国建国前出版的图书,最重要的工具书是《民国时期总书目》。查找近现代出版的丛书主要用《中国近代现代丛书目录》。查找近现代图书还需要注意利用一些大的出版机构的图书目录,因为这些出版机构出版数量大、所出版的图书影响也大,如《商务印书馆图书目录》、《中华书局图书目录》、《生活·读书·新知三联书店图书总目:1932—1994》等。

查找中华人民共和国成立以后出版的图书,主要用《全国总书目》、《中国国家书目》、《全国新书目》、《新华书目报》等。《全国内部发行图书总目》可与《全国总书目》配套使用。

查找已经出版的图书,除了使用上述目录工具外,还可使用大型图书馆的馆藏目录,如公共图书馆、科学院图书馆及高校图书馆的馆藏目录。这些馆藏目录一般都会建立在这些机

构的数字图书馆中,为用户提供网络检索的便利。如国家图书馆(<http://www.nlc.gov.cn>)的馆藏目录(<http://opac.nlc.gov.cn>)就提供了这个国内最大的文献收藏单位的丰富的馆藏信息,给用户提供了网络检索的条件。并且,国家图书馆同时兼具我国版本图书馆的功能,故利用该馆的馆藏目录也能够保证一定程度的查全率。该馆联机公共检索目录高级检索页面如图7.1所示。人们利用国家图书馆的目录,除了了解国家图书馆本身的馆藏外,还可以利用国家图书馆的版本图书馆的性质和该目录系统提供的网络检索的条件,来了解国内有关特定主题图书的出版和收藏的情况。

图 7.1 国家图书馆联机公共检索目录(OPAC)高级检索页面

中科院国家科学图书馆在科技文献信息的收藏、管理与利用方面发挥着重要的作用。在中科院国家科学数字图书馆的目录体系中,首先可以提供查找国家科学图书馆的馆藏资源,还提供该馆与中科院系统中所有科研院所的馆藏资源,包括一些特藏资源。人们可以通过这个数字图书馆的馆际互借和文献传递系统进行检索。另外,该系统中还有一个跨越全国的期刊联合目录系统,覆盖了全国范围内数百个图书馆的期刊馆藏资源,为人们查找科技期刊文献的原文提供了可靠的保证。图7.2是这一目录的首页。

图 7.2 中国科学院国家科学数字图书馆全国期刊联合目录首页

中国高等教育文献保障系统(CALIS, China Academic Library & Information System)是经国务院批准的我国高校信息资源共建、共知、共享的一个公共服务体系。其中的联合目录数据库子系统,覆盖了全国数百所高校的印刷型图书和连续出版物、电子期刊和古籍等多种

文献类型；覆盖了中文、西文和日文等语种；内容囊括了教育部颁发的关于高校学科建设的全部的二级学科、三级学科，是人们查找馆藏文献的重要工具。图 7.3 是 CALIS 联合目录公共检索系统的高级检索页面。

CALIS联合目录公共检索系统
CALIS Online Catalog

Simple Query Advance Query Ancient Books Scan Query History My Record List Authority Search Question Help Exit System

Advance Query You are searching in the center database of CALIS OPAC united catalog.

题名 包含 与 否
责任者 前方一致 与 否
主题 包含 中图分类号表

限制性检索 检索 重置

内容特征 所有 语种 所有 (Format: YYYY)
出版年份 不限
Material Type ☒ 图书 ☒ 连续出版物 ☒ 中文古籍 ☒ 多媒体 ☒ 电子资源 ☒ 视频资料
清除 全选

说明:

1. 请选择检索点，输入检索词，选择限定信息，点击“检索”按钮或直接回车；
2. 默认的检索匹配方式为前方一致，也可以在复选框中选择：精确匹配或包含；
3. 最多可输入三项检索词，默认逻辑运算方式为“与”，也可以在复选框中选择“或”、“非”；
4. 选择分类号检索点，可以点击“中图分类号表”按钮浏览，选中的分类号将自动填写到检索词输入框中；
5. 限制性检索的文献类型可选择：普通图书、连续出版物、中文古籍，默认为全部类型；
6. 限制性检索的内容特征可选择：统计资料、字典词典、百科全书，默认为全部；
7. 可通过输入出版年份对检索结果进行限定，例如：选择“介于”并输入“1999-2000”，即检索1999年至2000年出版的文献；
8. 检索词与限制性检索之间为“与”的关系；

图 7.3 CALIS 联合目录公共检索系统的高级检索页面

而其他一些图书馆，尤其是专业、专题的图书馆的馆藏目录对于查找特定的图书、特藏的文献也有着不可或缺的作用。

2. 外文目录举要

1) 检索外文书刊出版信息的目录

(1) 美国在版书目 (BIP, Books In Print)。

美国的出版信息多通过出版商目录来收集。其中，美国鲍克公司 (<http://www.bowker.com>) 的数据库 Bowker's Books in Print Database 是一个重要的工具。该目录最早源自鲍克公司《美国出版商目录年报》(PTLA, Publishers Trade List Annual)，这是美国出版商、发行商的在版书目总集，集中收录了美国出版商提供的书目，除将各出版商送交的原始书目的书型、尺寸加以统一并加上少量的出版索引外，其余一律照旧，不做更改。因此，其编制体例、著录格式和详尽程度各不相同。1948 年，鲍克公司开始印行《美国在版书目：著者、书名》(Books in Print: Author, Title) 目录，提供从书名与著者角度的检索途径。1957 年起又编制发行《美国在版书目主题指南》(Subject Guide to Books in Print)，又提供了主题角度的检索途径。这两种书目专门收录美国出版的或在美国发行的英语图书，提供了多个角度检索已出版图书的入口，解决了 PTLA 不便于检索图书的问题。BIP 系列的书目是以赢利为目的的综合性书商书目，只登记和通报在市场上销售的书，不作为推荐性的目录。

Bowker's Books in Print database (<http://www.booksinprint.com>) 是这一系列书目的电子版，目前在网络上提供服务。网络版 BIP 除了收录印刷版的同名书目外，还包括下列出版物或数据库：Books Out of Print (绝版书目)、Children's Books in Print (儿童书目)、Forthcoming Books (近期出版物书目)、Bowker's Publisher Authority Database (Bowker 出版商数据库)、Bowker's Complete Video Directory (Bowker 视频出版物指南)、Ulrich's Periodicals Directory (乌利希期刊指南) 等。近年来，该数据库同时发行全球版书目 (www.globalbooksinprint.com)，在原先收录美加图书出版信息的基础上增添了其他一些英语国家的出版信息，包括英国、澳

大利亚、南非和新西兰等国在内的英语和西班牙语的图书，形成了全球最大的在线图书书目检索系统，包含在版、绝版和即将出版的近数千万种图书和音像制品的书目信息。全球版书目的检索界面与功能与 BIP 美加版相同。这是个收费数据库。

与印本书相比，网络版的 BIP 增加了更多的功能，如该系统整合多种书商目录；提供多种检索途径；链接本地图书馆目录（Hooks to Holdings）；链接网上书店，提供检索途径以查找价格最便宜的图书；可以形成用户自己的书目记录，保存检索记录后可得到类似的新书预告；详细的出版商信息（历史沿革、母子公司、联系信息等）；详细的书评信息；作者简介与链接；书奖信息，畅销书信息；系统可根据用户的检索操作形成购书单；第 1 章试读等。图 7.4 是网络版 BIP 的快速检索界面。



图 7.4 网络版 BIP 的快速检索界面

(2) 英国国家书目 (BNB, British National Bibliography)。

这是英国图书馆书目服务部以英国图书馆缴存本图书馆所收到的图书作为编目基础，及时报道英国和爱尔兰出版和发行的图书和新版期刊的书目，这些文献资源的数据始自 1950 年。英国的缴存本图书馆除英国图书馆外还有多家，因此，从 20 世纪 90 年代后，该目录扩展到所有的缴存本图书馆的文献，随着英国缴存本范围的扩大，自 2003 年起，该目录又逐渐扩展到电子出版物。除了收录已出版的图书等文献外，该目录系统还收录即将出版的新书信息。英国国家书目的出版发行格式包括每周的马克格式交换文档、每周的印本目录和每月的 CD 目录（CD-ROM 格式的目录自 2008 年年底取消）。目前该目录可以在互联网上检索，具体网址为 <http://bnb.bl.uk/>，用户可以免费检索其简单的书目信息，但详细内容目前只有注册用户才可以获得。

(3) Ulrich 期刊指南 (Ulrich's Periodicals Directory)。

Ulrich 国际期刊指南是全球范围内的期刊名录数据库，也是美国鲍克公司的目录产品，创刊于 1932 年，最初由美国纽约公共图书馆期刊部主任 Carolyn Farquhar Ulrich 创始。原名为《乌利希国际期刊指南》(Ulrich's International Periodicals Directory)，现名为 Ulrich's Periodicals Directory。目前该目录同时以网络数据库 (<http://www.ulrichsweb.com/ulrichsweb>)

的形式提供服务。该数据库提供了世界范围的多于 300 000 种正式出版的期刊和非正式的连续出版物,既包括学术性的期刊,又收录开放存取的出版物;既有经同行评议而选出的刊物,又有大众化普及性的杂志;既有收费刊物,又有免费刊物,还有一定数量的报纸、快报。该系统按照杜威十进分类法来确定所收录刊物的学科类目,按照美国国会图书馆主题词表来标引所收录刊物的主题,共包括 100 多个族首级主题词。网络版的 Ulrich 期刊指南除了保留原有印刷版的所有功能外,还提供了期刊全文文章的链接等服务。这个数据库的检索方式包括快速检索与高级检索两种,可以按照 ISSN、关键词、学科主题、完整刊名、刊名中的关键词等快速查找,也可以按照学科主题、ISSN 或 CODEN 码、出版国别、语种、分类号、电子版提供商等多种方式浏览。该数据库多用于查找与期刊出版有关的各类问题,如期刊刊名的变更情况、期刊被文摘索引数据库收录的情况等。

2012 年年底出版的第 51 版纸本目录,4 大卷,1 975 美元,ISBN 号为 978-1-60030-638-9。印刷本第 51 版是有关期刊及连续出版物方面的参考咨询、研究和管理的基本参考工具书,提供了最新的国际期刊、报纸的书目信息,收录 200 多个国家和地区、9000 多家出版社的 22 万多种期刊及连续出版物,按 903 个主题类排列,是目前世界上最全的期刊目录,包括 10 000 种新刊及连续出版物、3434 种停刊、光盘版期刊、网络版期刊等。每种期刊及连续出版物提供详细书目信息,有几十项之多,如题名、ISSN、创刊年、订费、刊别、出版社及联络资料、出版国别、内容语文、是否持续出版、期刊性质、过刊采购信息、分类号、广告刊登价格、主题范围、期刊内容摘要、被哪些索引及文摘收录等。书后有丰富的辅助索引。除出版印刷本外,还有网络版(<http://www.ulrichsweb.com/ulrichsweb/>),网络版信息更新更快,检索更方便,可从分类、关键词、订购率、出版者名称、题名、编辑名称等途径检索。尽管覆盖面很大,但该目录收录范围一直以来以美国等西方国家期刊及连续出版物为主,但最近几年收录中国、日本等亚洲国家出版物的范围有所扩大。

2) 检索外文书刊收藏信息的目录

(1) 美国全国联合目录(NUC, National Union Catalog)。

这是以美国国会图书馆馆藏为主体的北美最大的书本式联合目录。

美国全国联合目录的历史可以追溯到 20 世纪初。1900 年,美国国会图书馆馆长 Herbert Putnam 采用卡片目录,开始编制美国全国联合目录。1942 年以前,美国国会图书馆以寄存方式,将卡片目录分赠各大图书馆。在长期的卡片目录基础上,1942 年美国国会图书馆出版“A Catalog of Books Represented by Library of Congress Printed Cards”,内容是 1898—1942.7.31 的卡片,共有 167 册;1953 年改名为“National Union Catalog”。1956 年改为以月刊、季刊、年刊的方式发行。1956 年后扩大到北美 1100 多个图书馆的收藏,是世界上迄今卷数最多的书本式目录。由于国会图书馆收藏范围的世界性,NUC 被称为“世界书目”。

(2) 美、加连续出版物联合目录(Union List of Serials in Libraries of the United States and Canada)。

这是收录北美连续出版物的联合目录,最初的印本书收录美国和加拿大 956 个图书馆收藏的 1950 年以前出版的 156 499 种连续出版物。1950 年后更名为 New Serial Titles (新连续出版物联合目录)。后来又发展成为(期刊联合目录数据库) Union Lists of Periodicals,在美国计算机联机中心 OCLC 中提供服务。该数据库包括数千种期刊的馆藏情况,每一条记录中都列出了 OCLC 的成员馆收藏这种期刊的每期的情况。

(3) OCLC 世界书目(WorldCat)。

WorldCat 是世界范围图书馆的图书和其他资料的联合目录数据库,是美国联机图书馆中心(OCLC, Online Computer Library Center)系统中的核心数据库之一。OCLC 始建于 1971 年,当时只联合了美国俄亥俄州的 50 多个大学和学院图书馆的联机编目中心(早期 OCLC 的

含义是 Ohio College Library Center)。随着计算机和互联网的发展,该系统在联机目录领域快速发展,其中的 WorldCat 目前已成为连接世界范围内 110 多个国家或地区的 400 多种语言或方言的 6 万多个图书馆的最大的提供馆藏信息的数据库之一。该系统的每条记录中都带有馆藏地点,它包括以下类型的目录资料:图书、手稿、计算机数据文件、地图、计算机程序、乐谱、影片和胶片、报纸、期刊、录音资料、视频资料、网络资源等。除了报道一般的常规书目数据外,还包括其他一些有价值的信息,如内容目录、封面设计艺术、内容提要、著者简介等。提供的书目信息最早始自公元前 1000 年,从泥版书到电子书、从早期的唱片到如今的 MP3,从纸莎草纸手稿到网络地址,范围之大、内容品种之多,非任意一个独立的图书馆能与之比拟的。该数据库以前只在美国 OCLC 的 First Search 系统中提供服务。自 2005 年起(经过一年的试验后),OCLC 与 Google 和 Yahoo!这两大搜索引擎合作,通过搜索引擎也能检索到这个书目数据库的信息。

(4) 美国国会图书馆联机目录 (Library of Congress Online Catalog), <http://catalog.loc.gov/>。

这是反映美国国会图书馆基本馆藏的书目工具,除了反映馆藏的著录外,该数据库的描述还包括参见、注释、流通状态,还包括那些处于采购编目状态的文献。该数据库覆盖的文献类型包括图书、期刊及连续出版物、计算机文档、手稿、图谱、乐谱、有声文献、图像文献等。该数据库还提供 1984 年以来的中文、日文和朝鲜文的译文资料目录,以及 1988 年以来的希伯来语和依地语目录、1991 年以来的阿拉伯语和波斯语目录。该数据库提供基本检索 (Basic Search) 和引导检索 (Guided Search) 两种检索方法,并有语种、年代等限制条件。

7.2 索引与文摘

7.2.1 索引的概念

索引是将原始文献中某些重要的或有意义的信息,如书名、刊名、篇名、主题、人名、地名等分别择录出来,进行标引,再按一定方式编排,并注明出处,以供检索的工具。它包括四个基本要素:索引源、索引款目、编排方法和出处指引系统。索引也是一种传递文献信息、揭示与检索文献的工具,从这个意义上说,索引与目录有异曲同工之处。但索引与目录也有差别,目录是以文献整体为记录和检索单元的,如一册书、一种期刊、一次会议、一本论文集等(换句话说,目录是以单本文献为报道单元的),而索引则是以文献中的个别事项和内容作为记录和检索单元的,如一种期刊中的一篇文章、一本论文集中的一篇论文、一本书中的一个章节、一次会议中的一篇报告等(换句话说,索引是以单篇文献为报道单元的,并通过告知这篇文章所在的位置或出处,起到指南、导向、示址的作用。)

索引不仅为用户提供多种文献检索途径,并能通过检索词的使用,反映某一文献的主题内容及关于某一学科或课题的最新观点和发展趋势。同时,索引作为一种应用范围十分广泛的检索工具,在互联网的条件下繁荣发展,更成为人们全面、准确收集基础数据并进一步分析利用的最常用的情报工具之一。还需说明的是,在印本时代因其篇幅小的优点常区别于后文所要提及的“文摘”检索工具的“索引”,在电子本时代不再有篇幅压力的情况下,多与文摘合为一体呈现。

7.2.2 索引的类型

1. 按索引对象及挖掘的深度划分

1) 篇目索引

篇目索引是只标引索引对象(如图书或报刊文章)的篇目信息(只标引篇名、作者、出

处信息，不标注正文内容），按一定的编排顺序将所提取的标识组织起来并提供原文所在位置的检索工具。在国内，篇目索引有时也被称为“题录”。篇目索引的主要作用是查阅报纸、期刊、会议录中的文章。

2) 内容索引

内容索引是深入到文献的具体内容，提取出表征内容特征的字、词、句、主题等反映原文内容并具有检索意义的信息，按一定的编排顺序将所提取的标识组织起来，并提供出处的检索工具。内容索引一般多为附在专著或年鉴、百科全书等工具书之后，并按主题词、人名、地名、事件、概念等内容要项编排的书后索引。

2. 按索及对象的文献类型划分

1) 期刊索引 (Periodicals Index)

期刊索引就其所涉及的文献范围可分为专刊索引和多刊索引两类。专刊索引是某种期刊的辅助检索途径，一般都伴随该期刊连续出现，或置于每期现刊之末，或单独成册出版。前一种方式最为常见。许多外文学术期刊都有辅助索引，并多有不同周期的累积索引，如期索引、季度累积索引、半年累积索引、全年累积索引。一些历史久远的期刊还有跨年的累积索引，可以起到回溯性期刊文献检索工具的作用。所谓多刊索引是相对于专刊索引而言的，即收录多种期刊的索引。

期刊索引可按索及文献的时间分为现行索引和回溯性索引两类。现行期刊索引以当前期刊论文为收录对象，定期连续发行，并有年度累积本。最早出现的现行期刊索引是美国在 1879—2004 年间连续出版（除 1899—1902 年间有脱节外）的包罗世界医学期刊文献的 *Index Medicus*（《医学索引》）。回溯性索引是指以某一时期期刊文献为收录范围所编制的索引，如西方国家著名的回溯性和综合性期刊索引——*Poole's Index to Periodicals Literature* 就是查找 19 世纪期刊论文的最重要的工具。随着互联网的发展，许多数据库加大了对回溯数据的开发力度，如美国 Thomson-Routers 公司的 *Web of Science* 数据库，其原印本刊分别创刊于 1963 年的《科学引文索引》和 1972 年的《社会科学引文索引》的两大引文数据库，现网络版的均已回溯到 1900 年；我国上海图书馆的《全国报刊索引》，其印刷本创刊于 1955 年，而其网络版的数据已回溯到 19 世纪中期。回溯数据库在文科领域尤其兴盛，美国 H. W. Wilson 公司的多个数据子库都在大力推行回溯子库。

从索引内容的学科覆盖面来看，又有综合性期刊索引和专科性期刊索引之分。综合性期刊索引如《全国报刊索引》、*Readers' Guide to Periodical Literature* 等。专科性期刊索引所索及的学科范围则相对较狭，但内容集中且挖掘深入，如 *Humanities Index*, *Education Index*, *Alloy Index*, *Short Story Index*。

2) 报纸索引 (Newspaper Index)

报纸索引在其许多款目下有简明提要，因而具有文摘的作用，不仅可用于查明关于某事件或消息的来源，也可直接用于回答什么事情发生于何时何地或始末等问题。事实上，报纸索引既是报纸内容的检索工具，同时又兼备资料性工具书的作用。例如，*New York Times Index* 就是比较著名的外文报纸索引。

3) 文集索引 (Index to material in collections)

文集是一人或多人著述的汇编，如诗歌集、论文集、译文集、演讲集、会议录、某著者专集或选集等。文集以书的形式出版，在目录中反映的是该文集的整体情况——书名、编者、出版者、总页数等，反映不出文集中具体文章的内容。揭示文集内容的工具是文集索引。如果没有文集索引，散布在成千上万文集的具体文章是很难查找的。文集索引以人文科学、尤以文学类的居多，社会科学和科技的文集索引则较少。中文的文集索引在古籍中的应用较

多,常用的文集索引有《清代文集篇目分类索引》、《元人文集篇目分类索引》等。外文文集索引多用在文学类的作品方面,常用的有 Essay and General Literature Index (散文和一般文献索引)、The Columbia Granger's Index to Poetry in Anthologies (哥伦比亚格兰杰世界诗歌索引)、Short Story Index (短篇小说索引)、Play Index (戏剧索引)等,原先的印本索引现在都可以在 H.W.Wilson 数据库中检索。

4) 书评索引 (Book Review Index)

书评是研究和了解某书或某人的学术思想、政治观点或艺术风格的不可缺少的资料,是推荐图书、指导阅读的工具,也是评价和选购图书的重要依据之一。通过书评还可以了解书刊出版后的反映情况、社会效果、有关研究领域的发展动向及出版国当前在政治、社会、文学、艺术和科技方面的现状等。

书评往往刊载在专门的书评刊物、图书馆类杂志、各专业杂志或报纸的副刊上。把分散在各处的书评收集起来并指引出处,即为书评索引。也有一些索引工具书附有书评索引,如 Readers' Guide to Periodical Literature、Library Literature 都将正文中的书评索引款目提取出来,汇总成书评索引附在书后,方便用户检索书评信息。

5) 会议录索引 (Proceedings Index)

会议文献的称谓很多,如会议录、会报、议事录、会议文献汇编、学术报告集。英文常用 Transactions 表示会议上发表的论文,用 Proceedings 表示会议的记录和会后整理出版的会议文献。由于会议录汇集了会议上发表的论文、讲话和报告,成为获取科学技术信息和学术研究成果的重要来源。各种会议规模、类型、性质不同,有鉴定会、研讨会、座谈会、专题讨论会等(英文名称有 Congresses、Conferences、Conventions、Seminars、Symposia、Workshops 等),内容重叠交叉,有时也很难区分。会议录的形式也很复杂,有会前预印本、论文摘要、会议期间的论文汇编等,有的会议录作为图书、论文集、丛书出版,有的则以期刊特辑、声像资料的形式发行。

与其他出版物相比,会议录往往能更快地反映前沿科学的新动向和新成果,内容也较专深。有的会议录论文不再在其他出版物上刊载,难以获得,所以它普遍受到科研人员的重视。因而针对会议录的这些特点,要力求及时查明,尽早获得和使用。

6) 引文索引 (Citation Index)

利用论文后引用的参考文献去追溯另一批相关的文献,这样层层追踪,直至获得满意的结果,这就是引文法,其检索工作量非常大。但有了引文索引就大大方便了这种查找方式。引文索引是利用文献之间的相互引证关系来检索文献的。从引文索引中查出一批所需的文献后,再利用这些文献的引文查找一批新的文献,这样不仅能获得一定数量的相关文献,还能揭示旧文献对新文献的影响、新文献对旧文献的评价,展现新旧文献在学术研究中的关系。由于引文索引及各学科的文献,因而还能从不同角度揭示学科交叉和相互渗透的关系。另外,根据引用频率的高低,结合其他方法,能为评价某一论文、某一期刊、某一著者、某一机构、某一地区,甚至某一国家的学术水平和产生的社会影响提供相对比较客观的依据。

引文索引思想最早是在 1955 年由美国学者加菲尔德 (Eugene Garfield) 提出的。在他的主持下,美国费城科学信息所 (公司) (ISI, Institute for Scientific Information) 于 1961 年推出了《科学引文索引》(SCI, Science Citation Index), 1973 年出版了其姐妹篇《社会科学引文索引》(SSCI, Social Science Citation Index), 1978 年又出版了《艺术与人文科学引文索引》(A&HCI, Art & Humanities Citation Index)。随着文献信息的电子化和互联网的出现,引文索引又在印刷版的基础上相继发展了光盘版和网络版。

在深入了解、研究引文索引之前,必须先弄清以下几个有关概念。

(1) 原文 (来源文献): Citing Papers 或 Source Items, 列有参考文献的原始文献, 如各类论文、图书等。引文索引中的原文一般以期刊文献为主。

(2) 引文 (被引文献): Cited Papers、Cited Reference 或 Citations, 被来源文献著者引用的文献, 如论文或图书后列出的参考文献就是一种典型的引文。不过, 英文 Citation 一词用在其他文摘、索引类的检索工具中也有款目、条目的含义, 相当于英文中的 Entry 一词。在更狭义的含义上, Citation 有时仅表示一篇摘要。

(3) 来源著者: Citing Author, 引用文献的著者, 即原文或来源文献的著者。

(4) 被引著者: Cited Author, 被引文献的著者。

(5) 引文耦合: 若文献 A 引用或参考了文献甲, 则文献甲是文献 A 的引文; 而文献 A 提供了包括文献甲在内的若干引文, 则将文献 A 称为来源文献。若来源文献 A 和来源文献 B 都引用了引文甲, 则称文献 A 和文献 B 为引文耦合, 而文献甲就是它们的引文耦。引文耦愈多, 其相应的来源文献之间的相关性愈高。

(6) 同被引: 若有两篇引文文献甲、文献乙共同被后来的一篇或多篇文献所引用, 则称文献甲、文献乙之间有同被引关系。同被引频次 (或称同被引强度) 愈高, 则其相应的来源文献间的关系愈密切。

(7) 自引: 来源文献的著者引用自己先前发表的作品, 则被称为自引。自引一般可以反映某项或某些研究工作间的承接关系。

文献之间的引证关系反映了一种科学交流活动, 显示了科学文献之间、刊载文献的期刊之间及文献所属学科之间的内在联系。论文之间的互相引证和被引证的关系使许多论文联系起来, 构成论文网, 这些论文的著者也因此被引文联系了起来, 构成著者网, 并在相关的学科领域形成文献网。通过追溯文献之间被引文联系起来的这种引文关系, 可以找到一系列内容相关的文献, 据此编制成以被引用文献的著者为标目的索引, 开辟了一种检索途径。

通过参考文献 (Cited Reference), 人们可以发现某一作者的研究如何受到前人的影响; 通过被引用次数 (Time Cited), 人们可以了解一篇论文 (实际上也反映了一项研究) 对之后的研究所产生的影响; 通过相关文献 (Related Records), 人们可以很方便地检索到引用相同文献的论文。

20 世纪 90 年代后期, 我国已有综合性引文索引和专科性引文索引, 如中国科学院编制的《中国科学引文数据库》(CSCD)、南京大学编制的《中文社会科学引文索引》(CSSCI) 等。一些期刊全文数据库也编制了引文索引, 如同方、万方、维普的引文库。我国台湾也编制了《台湾社会科学引文索引》(TSSCI) 和《台湾人文学科引文索引》(THCI)。

3. 按索引语言划分

1) 题名索引

根据索及文献中的题名而编制的索引, 一般有书名索引、刊名索引、篇名索引等。

2) 著者索引

根据索及文献中的责任者而编制的索引, 一般有著者索引、专利发明人索引等。

3) 机构、团体索引

根据索及文献中的著者单位或团体著者而编制的索引, 一般有团体索引、机构索引及著者附属单位索引等。

4) 分类索引

根据索及文献中的分类号而编制的索引。图书、期刊、会议文献中都会有这分类索引, 国内多为按中图法编制的中图分类号索引。也有为专利文献而编制的国际专利分类索引。国外的分类体系较多, 名称也不尽相同。

5) 主题索引

按索及文献中的主题词而编制的索引, 一般有标题词索引、关键词索引等。

6) 引文索引

按索及文献的参考文献信息及索及文献被他人引用的信息而编制的索引, 也包括索及文献本身的基本信息。在印本索引工具中一般有引文索引(即按索及文献的引文而编制的索引, 如引文著者索引、引文期刊索引等)及来源索引(这是根据索及文献本身的内容而编制的索引, 如著者、机构、关键词等常用索引)。在机检数据库中, 引文数据库还提供索及文献被他人引用的文献信息(关于引文索引的详细内容可参见 7.2.2 节的 2 中按索及文献类型分类下的引文索引介绍)。

7) 号码索引

号码索引是根据索及文献的各类号码而编制的索引, 如期刊文献中的国际连续出版物号索引、化学文摘中的化学物质登记号索引、专利文献中的专利号索引、科技报告中的报告号索引、合同号索引、资助号索引、标准文献中的标准号索引等。

7.2.3 索引举要

索引是信息组织的主要成果之一, 也是信息收集环节的重要检索工具之一。从古到今, 从国内到国外, 无论是综合性领域, 还是专业性学科, 无论是印本书, 还是电子版、网络版, 索引无处不在。尤其是在网络数据库迅猛发展的今天, 许多网络索引工具都相继建立索引与全文间的链接, 使索引的应用更加广泛、索引的发展更加繁荣。有关索引的详细举要将会在专业的图书中介绍, 这里只列举几种最常用的、综合性的索引工具。

1. 中文索引举要

1) 《全国报刊索引》

上海市图书馆在 1955 年 3 月创刊了《全国主要期刊资料索引》, 初期为双月刊, 1956 年起改名为《全国主要报刊资料索引》, 同年下半年起改为月刊, 1959 年起分成《哲学社会科学版》与《自然科学技术版》两刊, 一直出版至 1966 年 9 月, “文革”开始后休刊。1973 年复刊时正式改名为《全国报刊索引》(月刊), 前期哲社版与科技版合一, 1980 年又分成《哲学社会科学版》与《自然科学技术版》两刊, 出版至今。在印本索引的基础上, 由文化部立项、上海图书馆承建开发了《全国报刊索引数据库》, 其收录的内容及收录范围与《全国报刊索引》基本相同, 但在数量与收录报刊品种上多于后者。该库集全文库、索引库、专题库、特色库、报纸库、会议论文库为一体, 最早的数据回溯到 1833 年, 如《晚清期刊全文数据库(1833—1911)》、《民国时期期刊全文数据库(1911—1949)》、《晚清期刊篇名数据库(1833—1911)》、《民国时期期刊篇名数据库(1911—1949)》、《近代民国中医药专题库》、《音乐戏剧戏曲专题库》、《全国报刊索引数据库——会议库》、《家谱数据库》以及全国报刊索引数据库——目次库》和《全国报刊索引数据库——篇名库》等。图 7.5 是《全国报刊索引》网络版的高级检索界面。

2) 中国科学引文数据库

中国科学引文数据库(CSCD, Chinese Science Citation Database)由中国科学院文献情报中心创建于 20 世纪 90 年代, 这是我国第一个引文数据库, 曾获中国科学院科技进步二等奖。1995 年 CSCD 出版了我国的第一本印刷本《中国科学引文索引》; 1998 年出版了我国第一张中国科学引文数据库检索光盘; 1999 年出版了基于 CSCD 和 SCI 数据, 利用文献计量学原理制作的《中国科学计量指标: 论文与引文统计》; 2003 年 CSCD 上网服务, 推出了网络

版；2005年CSCD出版了《中国科学计量指标：期刊引证报告》；2007年中国科学引文数据库与美国 Thomson-Reuters Scientific 合作，与 Web of Science 实现跨库检索。



图 7.5 《全国报刊索引》高级检索界面

CSCD 收录我国数学、物理、化学、天文学、地学、生物学、农林科学、医药卫生、工程技术、环境科学和管理科学等领域出版的中英文科技核心期刊和优秀期刊千余种，目前已积累从 1989 年到现在的论文记录 300 多万条、引文记录 1700 多万条。除具备一般的检索功能外，还提供新型的索引关系——引文索引。使用该功能，用户可迅速从数百万条引文中查询到某篇科技文献被引用的详细情况，还可以从一篇早期的重要文献或著者姓名入手，检索到一批近期发表的相关文献，对交叉学科和新学科的发展研究具有十分重要的参考价值。中国科学引文数据库还提供了数据链接机制，支持用户获取全文。

中国科学引文数据库分为核心库和扩展库，数据库的来源期刊每两年进行一次评选。核心库的来源期刊经过严格的评选，是各学科领域中具有权威性和代表性的核心期刊。扩展库的来源期刊经过大范围的遴选，是我国各学科领域的优秀期刊。中国科学引文数据库 2013 共遴选了 1143 种期刊，其中英文刊 126 种、中文刊 1017 种、核心库期刊 780 种、扩展库期刊 363 种。

3) 中文社会科学引文索引

由南京大学研制成功的“中文社会科学引文索引”(CSSCI, Chinese Social Sciences Citation Index)是国家、教育部重点课题攻关项目。1999 年南京大学中国社会科学研究评价中心开始研制并出版光盘版《中文社会科学引文索引》，同时提供网上查询与统计服务（包括网上服务）。CSSCI 遵循文献计量学规律，从中文人文社会科学学术性期刊中精选出学术性强、编辑规范的 714 种期刊（2012—2013）作为来源期刊，其中核心刊 535 种，扩展刊 179 种。该数据库主要分为“来源索引”和“引文索引”两部分。在“引文索引”中输入要查找的作者姓名、篇名等，即可得知道该作者或该论文被引用的次数，以及引用者的姓名、论文篇名及其出处。在“来源索引”中可以查找某一作者或某一论文引用其他文献的详细情况。索引的检索点很多，来源文献检索途径包括：篇名、作者、作者所在地区机构、刊名、关键词、文献分类号、学科类别、学位类别、基金类别及项目、期刊年代卷期等。被引文献的检索途径包括：被引文献、作者、篇名、刊名、出版年代、被引文献细节等。检索结果按不同检索途径进行发文信息或被引信息分析统计，部分数据可以链接全文。图 7.6 是中文社会科学引文索引的来源文献高级检索界面。

2. 外文索引举要

1) Web of Science

这是美国系列引文索引的合称。是一种以期刊论文、专利说明书、科技报告等类型的文献所附的参考文献的作者、题名、出处等项目按照引用与被引用的关系进行排列而编制的索引。引文索引思想最早是在1955年由美国学者加菲尔德(Eugene Garfield)提出的。在他的主持下,美国费城科学信息所(ISI, Institute for Scientific Information)于1961年推出了《科学引文索引》,1973年出版了其姐妹篇《社会科学引文索引》,1978年又出版了《艺术与人文科学引文索引》。随着文献信息的电子化和互联网的出现,引文索引又在印刷版的基础上相继发展了光盘版和网络版。

图 7.6 《中文社会科学引文索引》的高级检索界面

无论是印刷版还是光盘版、网络版,美国引文索引主要分为两大部分:来源索引部分和引文索引部分,其主要作用也可定义为文献检索工具和引文分析工具。

文献检索工具。与普通的文献检索工具相比,Web of Science的各个子库是经过影响因子等科学工具筛选后确定的世界范围内自然科学、社会科学和人文科学方面8000多种核心期刊的综合性的、多学科的、具有权威性的文摘类检索数据库。既可以从篇名、关键词、文摘、著者、来源期刊、机构、地址等角度检索世界范围的主题核心期刊文献信息,又可以查阅到在专业学科文献集合中难以查检的某些交叉学科的资料;与全文数据库相比,Web of Science具有收录期刊的学科范围广、地域覆盖面宽、时间跨度大等特点,可帮助用户全面地了解和掌握特定主题的研究成果的线索;与其他普通的水摘检索工具相比,引文索引还有参考文献的链接、被引文献的链接和相关文献的链接。

很多人将引文索引仅定义为引文分析工具,忽略了引文索引的文献检索功能,这实在是一个误区,也是使用这一数据库的一大损失。

引文分析工具。科学家认为,在一定的时期内一篇文章的重要程度或者说影响力可以通过被引用的程度来衡量。Web of Science独特地收集了每篇文章的参考文献,并且编制成索引。通过对引文的分析,可以了解某一学者的研究成果被他人关注的程度,以及对其他研究的影响。从某学科内部期刊、文献的引文所反映的主题相关性,可以了解某一学科的结构。从不同学科期刊、文献引用的网状和链状关系,可以揭示各学科之间的关系,可以展现某项目或事件的发生和发展,揭示某思想或方法的改善、扩充和修正,了解各领域的

前沿问题,从而得到完整的科学发展史,并预测未来的发展方向和热点问题。

(1) Web of Science 的收录范围。

Web of Science 是 ISI Web of Knowledge 系统中的引文索引数据库,由三种引文索引子库组成。其中,引文索引具体的文献覆盖年度分别为: Science Citation Index Expanded (1900—今)、Social Sciences Citation Index (1900—今)、Arts & Humanities Citation Index (1975—今)。Web of Science 引文索引部分收录世界范围内最有影响力的、经过同行专家评审、经过影响因子筛选的高质量的科技、社科、人文科学与艺术方面的期刊,其中 SCI 扩展版 (SCI Expanded) 收录了世界范围 150 多个学科的一流科技期刊 8500 多种; SSCI 收录 3000 多种社会科学类的核心期刊; A&HCI 收录 1700 多种人文社科类的核心期刊。由于学科交叉的原因,三个子库之间所覆盖的期刊也有一定程度的重叠。该数据库每周都会进行更新。

由于 Web of Science 的市场价格较高,因此买方一般不一定购买其全部数据库、全部回溯年代或全部链接的使用权,而是其中某一部分或某些部分。如国内一些单位只购买其中的 SCI Expanded,回溯年代则根据各订购单位的需要自定。图 7.7 是最新版 Web of Science 的主页面。

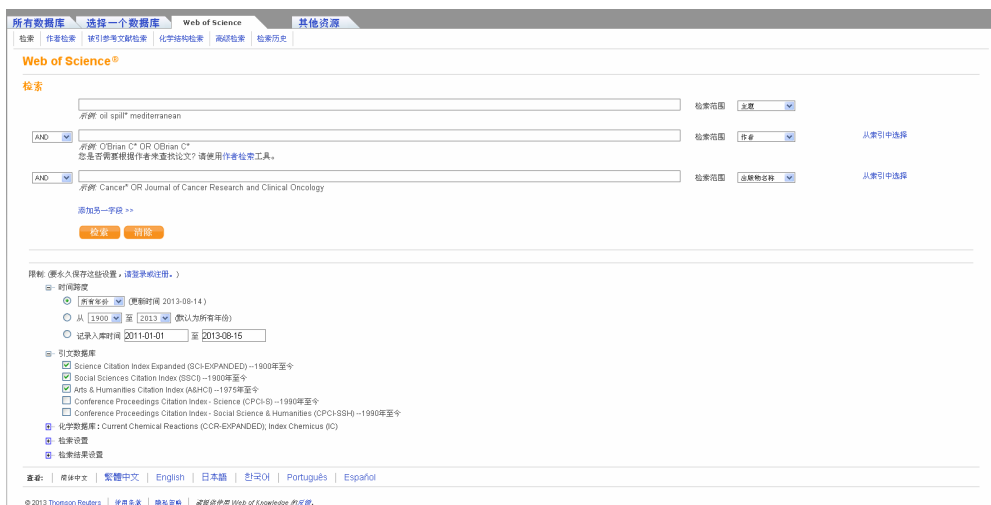


图 7.7 Web of Science 的主页面

(2) Web of Science 的特点。

网络版的引文索引除了同样具有与印刷版和光盘版的特点(如经过核心期刊的筛选、可以通过引文进行检索等)外,还具有提供跨学科检索、跨年代检索这一特点。由于印刷版和光盘版都受存储空间限制而只允许分年代、分学科(如自然科学、社会科学、人文科学分册或分盘收录)进行检索,而科学技术本身的发展却使学科之间的交叉、渗透与融合日益复杂。例如,公共健康、环境保护、计算机等原先看上去是自然科学领域的研究主题,现在却越来越多地被社会科学领域的学者所研究,而心理学、情报学、社会学等原先看上去是社会科学领域的研究主题也越来越多地被自然科学领域的学者所重视、所采纳。而 Web of Science 中全部的收录内容可以在同一个界面进行检索,使跨学科、跨年代的检索需求有了全面、准确获得检索结果的可能性。

(3) Web of Science 的检索功能。

Web of Science 数据库检索主页面主要的功能键包括:“检索”(Search)、“引文检索”(Cited Reference Search)和“高级检索”(Advanced Search)这三种选择。

“检索”是对来源文献的检索，包括主题内容（可以同时包括篇名、关键词和文摘三个字段，称为 TS 字段）、题名、著者、来源期刊、团体作者、期刊出版年和著者地址（著者地址检索可按机构名、部门名、街道名、城市名、国名甚至邮政编码检索）等检索字段。

“引文检索”的入口包括被引文著者、被引文献（如被引期刊、被引书名或被引专利号等）和被引文年代三项。

Web of Science 的“高级检索”有点类似于其他一些外文数据库中的“专家检索”（Expert Search），这个检索界面就是一个检索窗口，专门用于处理那些检索要求复杂、检索涉及项目众多的情况。所有的检索条件都要求输入检索字段的缩写符号，如题名字段（TI）、著者字段（AU）、来源期刊刊名字段（SO）等，要求检索者熟练掌握数据库的检索技巧，包括布尔运算等。

与绝大多数机检数据库一样，Web of Science 也提供逻辑与、或、非的检索，此外，该数据库检索组配符中还有一个运算符为 SAME，这是逻辑与运算（AND）的扩展，将该符号两边的检索词限制在同一个检索字段，如在 TS 字段输入 Internet SAME Search，系统操作时，只有当这两个检索词同时出现在篇名字段或同时出现在关键词字段或同时出现在文摘字段才算命中；但如果输入的是 Internet AND Search，那么，那些篇名中有 Internet、文摘中有 Search 的文献，也会被系统认为是匹配文献而被命中。

Web of Science 的截词检索允许右截断和中间截断，星号（*）为无限截断符，表示 0 到任意数的字符截断，如 SUL*UR*，可表示 sulfur、sulphur、sulphuric、sulphurous；问号（?）为有限截断符，一个“?”表示一个字符截断，两个“?”则表示两个字符截断。

从 2004 年暑期起，Web of Science 把对检索结果分析的功能提供给用户使用，这就是我们在检索结果页面上看到的“Analyze Results”。这里的分析主要依据在不同的选项中的文献量的降序排列来实现对检索到的数据集合的量化。单击“Analyze”（分析）键后，系统允许最大限度一次可分析数据 10 万条。分析的选项包括作者分析、国家或地区分析、文献类型分析、机构分析、语种分析、出版年分析、期刊分析、主题分析等。

需要说明的是，尽管数据库提供了很好的分析功能，但这一功能毕竟是通用性质的，对一些特别的分析要求并不能完全满足。例如，这里的著者分析指的是全部参与研究的作者，不分第一作者和其他参与的作者；主题分析也比较宽泛，特指度有限，用户在使用时还需根据特定的需要进行大量的人工操作以满足深入分析的要求。

Web of Science 原本是一个索引性质的、独立的数据库，但近年来开发商通过开发数据库的链接功能使这个数据库与其他一些数据库连接起来，大大增强了这一数据库的使用功能。如通过 Web of Science，研究人员可以链接到某些论文的全文，前提是如果图书馆同时订阅了该论文所在期刊的电子版本。该系统目前已与多个全文期刊出版商建立了全文检索连接关系。通过全文连接使这个索引文摘性质的数据库扩大了使用功能，帮助用户通过该数据库的文献线索直接看到原文。

2) Wilson 公司索引数据库

Wilson 网络数据库(<http://www.hwwilson.com>)是 HW Wilson 出版公司的产品。HW Wilson 的历史最早可追溯到成立于 1889 年美国明尼苏达大学学生宿舍的小书商。自从 1898 年“累积图书索引”（Cumulative Books Index）出版后，索引类的工具书开始成为该公司的发展方向。1901 年，“期刊文献读者指南”（Readers' Guide to periodical Literature）的问世赢得了很好的国际声誉；80 年代早期开始发行电子出版物；如今，该公司已有 64 种大型索引和目录数据库，并建立了基于互联网的数据库——WilsonWeb，该系统目前在 EBSCO 的平台上提供服务。

Wilson 网络数据库的范围包括应用科学技术、艺术、生物与农学、商务、教育、人文科学、社会科学等,其中许多数据库同时提供不同的文献版本,即索引数据库、文摘数据库、全文数据库,供用户根据自身的需求环境按需订购;也有在原先印刷本的索引工具书基础上形成的网络版,如读者期刊文献指南、论文与一般文献索引、商务期刊索引、图书馆学与情报科学索引等,但这些都在印刷本的基础上,借助网络的优势,做了大量的扩充,使之获得更好的检索效果和更便捷的链接;还有一些数据库的组合版本,以适用于不同的用户群;此外,还有大量的全文传记数据库、书目数据库。总的来说, Wilson 数据库偏重于社会科学、人文科学、经济商务及应用科学技术和综合科学技术。该公司主要的索引数据库如下所述。

(1) 期刊文献读者指南 (Readers' Guide to Periodical Literature)。

这是 Wilson 公司出版的一份历史悠久的期刊索引, 1900 年创刊。目前共收录发行量较大且“家喻户晓”的美国大众性通俗期刊 350 种左右,以政治、历史、时事、文体体育、保健等内容为主,但也涉及天文、宇航、物理、建筑、电影、戏剧等内容。印本书按著者和美国国会图书馆主题表的字顺统一编排,卷末附有按图书著者编排的书评索引。

网络上的该期刊索引有多个版本,包括:① 期刊文献读者指南全文版 (Readers' Guide Full Text, Mega Edition), 收录 215 种 1994 年以来的期刊的全文及 400 多种 1983 年以来的期刊索引,其中大量的文章含有摘要;② 期刊文献读者指南全文选择版 (Readers' Guide Full Text, Select Edition);③ 期刊文献读者指南回溯版 (Readers' Guide Retrospective: 1890—1982);④ 还有文摘版等。由于该索引集中了发行量最大和为公众所喜爱的杂志,所以是查阅期刊文章不可缺少的检索工具。

(2) 社会科学文献索引 (Social Sciences Index) 和人文科学索引 (Humanities Index)。

Social Sciences Index、Humanities Index 和前文介绍的 Readers' Guide 均为 Wilson 公司的期刊索引,但侧重不同。SSI 和 HI 所收录的大多是专业性和学术性较强的期刊,而 Readers' Guide 所收录的大多为综合性、普及性和流行的期刊;Readers' Guide 收录美国出版的期刊,而 SSI 和 HI 主要收录英语国家出版的期刊,不过仍然以美国为主。

其中, Social Sciences Index 在 Wilson 公司的系列产品中也有多个版本,包括:① Social Sciences Index 全文版,收录 215 种期刊的全文,时间回溯到 1995 年,有 625 种期刊的索引和摘要,其中约 400 种为同行评议期刊,时间回溯到 1983 年;② Social Sciences Index 回溯版,收录 1907—1983 年间的文献的索引和摘要;③ 也提供索引版和文摘版,以及印刷本。Humanities Index 的全文版收录 215 种期刊的全文,时间回溯到 1995 年,收录 600 种期刊的索引和摘要,时间回溯到 1984 年,其中 400 多种为同行评议期刊;Humanities Index 回溯版收录 1907—1984 年间的文献的索引和摘要。

上述索引在时间、地区和内容上既有分工,又有衔接和联系,有效地利用 Wilson 的这套索引系列需要特别注意这些区别及相互间的衔接和互补关系。

(3) 杂文和一般论文集索引 (Essay and General Literature Index)。

该索引每期收选美国、加拿大、英国约 300 余卷文集, 20 多种年度出版物和连续出版物,重点以社会科学、人文科学,尤以文学评论为主,也涉及经济、政治、历史等。其中,传记和人物研究的资料特别丰富,常被作为传记索引使用。目前已收录 7000 多种文集集中的约 86 000 篇短文或论文。时间自 1985 年至今;另有回溯本覆盖 1900—1984 年间的短文和论文 249 000 余篇。印本的索引半年出版一次,有 1 年、5 年和 7 年累积本。用户可以根据文集的编者或主题等检索到文集的信息,也可以根据文章的作者或主题等检索到文章及文章所在文集的信息,对于文学评论性质的文章,还可以通过被评书的作者或书名等入口检索到有关文集的信息。

(4) 艺术索引 (Art Index)。

该索引收录国际方面的艺术类期刊,包括英语、法语、德语、意大利语、西班牙语、荷兰语的期刊,地域范围也扩展到加拿大、拉美、亚洲及其他国家的艺术品、艺术家、艺术评论、展品评论方面的文献。该数据库收录 480 种艺术类期刊的文摘信息,其中许多是同行评议期刊,有 160 种期刊提供全文,全文期刊回溯到 1997 年。该数据库覆盖的学科范围包括广告艺术、古董、考古、建筑与建筑史、艺术史、当代艺术、服装、工艺品、装饰艺术、绘画、工业设计、内部装修、园林设计、电影、博物馆学、非西方艺术、油画、摄影、陶瓷、雕塑、电视、纺织、录像等。该数据库还提供 1929—1984 年间的回溯版本、印刷版本。

(5) 人物传记索引 (Biography Index)。

许多原始的人物传记文章最初会刊登在期刊或报纸上。其中有些文章可能会被收录在某些期刊索引或报纸索引中,但并不是以传记人物为检索目的的,因此,如果没有专门的检索工具,这些期刊、报纸上的人物信息就是分散的、无序的。而 Biography Index 就是以这些分散的人物信息作为收集对象的。该数据库是由同名的印刷版索引发展起来的,收录 1984 年至当前的人物信息,包括 Wilson 公司其他数据库中分散在 3000 多种期刊中人物信息,其他期刊的人物信息,还包括每年 2000 多种近期的图书及人物文集中的人物信息。除传记、自传外,还收录人物专访、讣告、信件、日记、文集、青少年读物、书评、书目、展览评论等。该数据库的最早的数据可以追溯到 1946 年,这是 Biography Index 的创建年。另外,还有一个人物信息更多的数据库,名为 Biography Reference Rank,合并了 Wilson 公司的多个数据库中的人物信息,包括作家信息、人物图像信息等。

(6) 书评摘要 (Book Review Digest Plus)。

该数据库目前收录有关于 66 万多种书的 1200 万份书评摘要。其中 115 000 篇书评有全文。这些书评信息来自数千种期刊,有数十种 Wilson 现有的数据库中的书评信息成为该数据库的信息源,其中有前文所介绍的多个数据库为其提供书评信息,如“期刊文献读者指南”、“社会科学索引”、“人文科学索引”,此外还有“商务期刊索引”、“法律期刊索引”、“教育期刊索引”、“图书馆学情报学期刊索引”等。目前的书评信息回溯到 1903 年。

Wilson 公司的系统中还有很多专业的数据库,包括应用科学技术、商务期刊、法律期刊、教育期刊、图书情报学期刊等,这里不一一列举。总的来说,Wilson 公司的索引数据库以人文社科领域的为主,也涉及少量的综合性的科技信息。

3) 美国信息集团科技文摘数据库

美国剑桥科学文摘社 (CSA, Cambridge Scientific Abstracts), <http://www.csa.com/>, 总部设在美国马里兰,是一个出版科学技术文摘索引类文献的私营信息公司。2007 年 CSA 与美国 ProQuest Information and Learning 合并。现拥有近百个网络数据库。覆盖的学科范围包括:生命科学、水科学与海洋学、环境科学、计算机科学、材料科学、航空航天及社会科学、人文科学等诸多领域。这些数据库大多是从传统手工检索工具延伸而来的,故这些源数据库的标引加工一般比较科学、专业、稳定。该数据库的检索结果为文献的题录及文摘信息。其学科范围涉及生命科学、水科学、环境科学、计算机科学、材料科学和社会科学等。具体的数据库(其中有的数据库在一个以上的学科中重复反映)如下所述。

(1) 艺术与人文科学部分的数据库有:现代艺术书目;Avery 建筑期刊索引;不列颠人文科学索引;艺术史书目;设计与应用艺术索引;设计师与设计公司文档;法国人文社会科学索引;伊斯兰索引;CSA 语言学与语言行为摘要;现代语言协会国际书目;哲学家索引;音乐文献摘要;等等。

(2) 自然科学部分的数据库有:农业索引;水科学索引;水科学与渔业文摘;生物学系

列数据库(包括生物学、水科学、海洋生物学,藻类、真菌、原生动物学、动物行为、细菌、钙和钙化组织、化学反应、神经学、生态学、昆虫学、遗传学、健康与安全、人类基因、免疫学、工业与应用微生物、核酸、致癌基因、生长因素、毒物学、滤过性微生物学和艾滋病方面的专题数据库;还有多个生物摘要方面的数据库);生物技术与工程文摘;生物技术研究文摘;会议索引;环境影响摘要;环境科学与污染管理;地球科学参考数据库;地球科学数据库;国际制药文摘;医学文摘;气象及地球天文物理文摘,美国科技信息服务局数据库;体育教育索引;植物索引;科技摘要;Scopus 自然科学部分;毒物学在线;水资源文摘;动物学记录;等等。

(3) 社会科学部分的数据库有:老年学数据库;应用科学索引与文摘;人际交流障碍文献;人际交流文摘;犯罪审判文摘、经济学文摘、教育学文摘、社会科学国际书目;国际管理机构数据库,伊斯兰索引,图书馆与情报学文摘;国家犯罪审判参考服务文摘;以及一系列的心理学方面的数据库;一系列的社会工作数据库,社会学数据库,政治学数据库等,还有 Scopus 商务与经济部分等。

(4) 工程技术部分的数据库有:航空宇宙与高科技数据库;新技术与工程文摘;水科学;生物技术与生物工程;生物研究文摘;微电子学与包装材料数据库;热物理数据库;国际计算机文摘;土木工程文摘;工程技术图表图像数据库,技术研究系列数据库;地震工程数据库;信息技术个案数据库;国际信息研究关注数据库;机械与交通工程文摘;纸业文摘;聚合物文摘等。

这些数据库的检索平台名为 CSA Illumina。

4) CALIS 外文期刊网(CCC, CALIS Current Contents Of Western Journals)

CALIS 外文期刊网是中国高校文献保障系统面向全国高校广大师生的一个外文期刊综合服务平台。它是普通用户获取外文期刊论文的最佳途径;也是图书馆馆际互借的基础数据来源;又是图书馆馆员进行期刊管理的免费使用平台。该资源收录 10 万余种国内高校收藏的纸本期刊和电子期刊,其中有 4 万多种期刊的文章篇名信息每周更新,目前期刊文章的篇名目次信息量达 8000 多万条。

7.2.4 文摘

国家标准《文摘编写规则》(GB 6447-86)将文摘定义为“以提供文献内容更改为目的,不加评价和补充解释,简明、确切地记述文献重要内容的短文。”

作为检索工具,文摘以简练的形式将文献的主要内容扼要地作成摘要并按一定著录规则与排列方式系统地排列起来。它不仅记录文献的基本书目信息,而且提供文献的内容梗概,是系统报道、积累和检索文献的重要工具。

文摘与索引一般都是以单篇文章作为报道单元的,实际上,文摘是在索引的基础上加上了文献主要内容的摘要。在印本书时代,文摘与索引还是有很大的区别的,例如,索引以报道文献的外部特征为主,而文摘以报道文献的内部内容为主;索引多为综合性的,而文摘多为专科性的;索引多数按主题或著者来编排,而文摘多数则按学科分类来编排;文摘的时差比索引的大;文摘所报道的文献的数量相对比较少。所有这些区别,多数是因为有了文摘后,印本书的篇幅比较大,在同单元、同大小的印本中,文摘可容纳的篇幅比索引要小得多。但在电子书本的时代,由于计算机的大容量,使得人们在文献加工时完全可以忽略篇幅的问题,加上期刊的电子化出版,也减少了数据库在文摘加工阶段的耗时,所以,在网络时代,从检索工具这个角度来看,索引与文摘的区别越来越小。许多印本时代的索引目前都逐渐发展为文摘性的工具,包括前文介绍的引文索引等。

7.3 全文数据库

7.3.1 全文数据库概述

1. 全文数据库的定义

全文数据库 (Full-text Database) 是收录有原始文献全文并能提供全文检索的一类文献信息数据库。与传统书目数据库相比, 全文数据库集文献检索与原文获取于一体, 是近年来发展迅猛、使用频率较高的一类文献数据库。

20 世纪 80 年代中期以来, 全文数据库在数据库中所占比例逐年增大。特别是进入 20 世纪 90 年代, 随着互联网的普及和网络技术的发展, 人们可以越来越方便和频繁地在网上访问和获取数字信息资源, 从而也就越来越希望在网上直接得到一次文献。因此, 全文数据库, 特别是基于互联网开发的全文数据库及全文服务飞速发展, 已成为文献数据库的主流。

2. 全文数据库的类型

全文数据库按照不同的标准可划分为不同类型。例如, 按照文件存储形式可划分为文本型、图像型和多媒体型; 按照收录的内容和学科专业范围不同可划分为综合型、专业型和专题型; 按存储服务形式可划分为光盘版和网络版等。通常, 对全文数据库的分类更多是依据其处理的常用对象类型来划分的。

(1) 报刊全文数据库: 以收录报纸或期刊上的文章为主, 如美国计算机学会的 ACM Digital Library 全文数据库、Emerald 全文数据库、中国知网学术期刊全文数据库、中国人民大学书报资料中心编选的《复印报刊资料》全文数据库、重庆维普的中文科技期刊全文数据库等。

(2) 商业信息、统计报告全文数据库: 收录各类市场新闻、公司信息、研究报告、调研报告等, 如 EBSCO 的“商业资源电子文献库”。

(3) 法律法规和案例数据库: 主要收录法律法规、司法解释、法院案例及判决书等, 如 LEXIS-NEXIS 的 LEXIS 数据库、万方中国法律法规全文数据库。

(4) 政府报告和新闻信息全文数据库: 如 LEXIS-NEXIS 中的 NEXIS 数据库。

(5) 其他对象类型的数据库: 如电子图书全文数据库、古籍全文数据库、专利全文数据库、标准全文数据库等。

3. 全文数据库的特点

与文摘数据库、索引数据库、引文数据库等文献数据库相比, 全文数据库具有以下特点。

(1) 信息内容的原始性。全文数据库集题录、文摘、全文于一体, 用户可以直接检索并获取原始文献。

(2) 信息检索的彻底性。除一般检索方法外, 还增加了全文检索和引文检索, 可以对文中任何字词句进行检索。

(3) 检索语言的自然性。既支持题录信息的规范语言检索, 又支持题录和全文的自然语言检索。

(4) 标引方式的灵活性。全文数据库一般都经过标引, 或者采用人工标引或者全文自动标引。

(5) 检索结果的准确性。全文数据库收录信息一般经过人工筛选, 信息质量高。

(6) 存储空间大。因为要存储各种文件格式 (PDF、CAJ、VIP 等) 的原文信息, 甚至多媒体文件, 所以全文数据库的存储空间较大, 通常达到 TG 级。

4. 全文数据库的评价标准

随着全文数据库应用越来越普及，对全文数据库的评价也引起学术界和信息机构的关注，建立全文数据库的评价标准，对于信息机构选择全文数据库、建设全文数据库资源和提供服务都具有重要意义。

从目前的研究来看，全文数据库的评价主要从资源内容和检索系统两方面衡量，采用层次分析法，可细分出如表 7.1 所示的评价指标体系。

(1) 学科情况。即全文数据库中提供全文的出版物学科领域覆盖范围，不同的全文数据库收录的学科范围和侧重点有所不同，数据库覆盖范围的广度、提供内容的多少是评价全文数据库资源的重要标准。

(2) 全文收录情况。所谓全文数据库，并非其收录的报刊、报告全部都是全文，因此全文占多大比例就很重要，通常全文占到 50% 左右即可以称为全文数据库，而全文能到 65% 以上就是比较好的全文数据库了。另外，全文是指整篇文献的收录，有些数据库号称收录全文，但实际上有时一篇文章由于版权等问题，只给出一部分篇幅，这样的“全文”是有水分的。

表 7.1 全文数据库评价指标体系

一级指标	二级指标	三级指标	说明
资源内容	权威性	学科情况	提供全文的出版物学科覆盖情况
		全文收录情况	收录全文的比例
		核心出版物收录情况	核心出版物占整个收录出版物的比例
		注销出版物收录情况	注销出版物全文收录情况
		资源加工质量	原文加工情况、题录质量、是否标引等
	时效性	覆盖时间	出版物收录全文的时间范围
		更新频率	全文数据库更新周期
		滞后周期	与印刷型出版物出版相比的滞后时间
检索系统	响应性	响应时间	检索响应的速度
	科学性	检索性能	包括检索技术、检索方式、检索途径
		检索结果	包括检索结果的排序、相关度、显示方式等
	易用性	检索界面	界面的友好性，是否包括导航、检索提示等
		原文获得	是否直接下载、共享使用

(3) 核心出版物收录情况。全文数据库往往为了求“全”而不断增加来源出版物，但全文数据库中核心期刊和权威出版物所占比例是说明数据库质量的一个重要因素。通常数据库中所包含的出版物品种的数量不足以说明问题，因为其中可能有不少是非核心刊物或通俗读物，这部分内容所占比重过大会大大降低数据库的质量和权威性。

(4) 注销出版物收录情况。全文数据库的出版物收录情况往往会发生变化，有些原来提供全文的出版物，虽然现在仍然包含在数据库中，但已经不再出版或停止向数据库商提供电子版或只提供文摘不再提供全文，即这些出版物并不包含当前的全文信息，这样就无法满足用户对最新数据的使用需求。因此要分析从数据库中注销的出版物的情况，如果过多，则数据库质量就会有所下降。

(5) 资源加工质量。全文数据库中全文信息主要来自印刷出版物的数字化加工和出版社全文电子版的提供，尤其是早期资源，多依赖于数字化加工，扫描、识别、校对质量高低直接影响全文数据库的原文质量。是否对原文进行描述、标引，生成高质量的题录也是评判全文数据库资源加工质量的一个指标，更是影响全文数据库检索性能的主要因素。

(6) 全文收录覆盖时间。早期出版物以印刷型为主, 文献数字化的浪潮出现在 20 世纪 90 年代末和 21 世纪初, 许多全文数据库加工一般是从 20 世纪 90 年代开始的, 因此, 全文数据库中全文收录时间也是判断全文数据库好坏的重要依据。

(7) 更新频率与滞后周期。数据库的更新频率越高, 内容的时效性越强, 通常以日更新或周更新为最佳。但由于全文数据库收录的内容仍以印刷型出版物为主, 也就存在着时滞, 即出版物被收录进数据库的时间与印刷型出版物的出版时间之间的时间差。时滞过长, 影响数据库的时效性和质量。有些全文数据库虽然收录的全文出版物很多, 但出版商出于版权的考虑, 限制全文上网的时间, 这些全文出版物最初只能提供给读者文摘或部分全文, 影响了读者查阅。如果这种出版物占比例过大或时滞过长, 数据库质量也相应下降。由于目前网络出版和数字出版的盛行, 越来越多的全文数据库甚至会优先于纸质期刊发布。

(8) 数据库检索系统。主要从响应性、科学性和易用性方面对系统响应时间、检索方式、检索途径、检索结果、用户界面等方面进行评价。其中, 能否进行全文检索, 能否在检索结果中用全文限定二次检索, 是否允许用户按相关性、日期或字母顺序重新排列结果, 能否原汁原味地展现全文 (包括图表信息), 是否提供打印、存盘或 E-mail 发送功能, 等等, 是评价全文数据库检索系统的一些主要指标。

目前, 对用户和服务的关注日盛, 越来越多的人开始重视全文数据库的服务评价, 把对数据库供应商的评价、用户培训与否、用户使用体验是否良好、是否提供个性化服务等也作为评价全文数据库的标准。

7.3.2 全文数据库开发

全文数据库既是一种信息资源服务方式, 又是一种资源保存方式。自 1973 年, 美国米德公司建成了世界上第一个面向公众查询的大型全文数据库 Lexis, 标志着全文数据库建设的兴起。随着计算机在信息产业的广泛应用, 20 世纪 80 年代中期开始, 国外全文数据库的建设呈现迅猛发展的势头, 我国则由于汉字处理的复杂性, 全文数据库的发展滞后一些, 20 世纪 90 年代中期以前, 只有一些大学和研究机构建成一些小型的全文数据库, 如武汉大学的《湖北省地方志大事记》、《中国人民解放军大事记》全文数据库, 陕西省中医研究院的中医经典古籍全文数据库、北京大学的中国对外经济贸易法律全文数据库等。到 20 世纪 90 年代中期, 全文数据库建设开始公司化运作, 形成了清华同方、北京万方和重庆维普三大全文数据库供应商及超星、书生之家、方正 Apabi 等电子图书供应商共同繁荣的局面。

1. 全文数据库的开发步骤

全文数据库开发包括全文数据库资源建设和全文检索系统建设, 本节主要阐述全文数据库资源建设的步骤。

全文数据库的资源建设步骤主要包括数据准备、文本预处理、数据加载、数据检索和数据维护几个环节, 图 7.8 展示了全文数据库的建库流程。

1) 数据准备

数据准备是指对计划加载到全文数据库中的数据收集、整理、归类等预处理的过程。加载到全文数据库中的数据可以通过多种途径获得, 常见的数据来源有 Office 文档、电子印刷产生的文稿、计算机网上传送的文件、电子出版物、图文处理产生的文件、专门组织人力录入等。数据收集起来之后, 要进行一些简单的分类, 一般按照数据内容进行分类, 同一类内容加载到同一库中, 这样便于查找。如果数据总量不大, 也可不进行分类。分类对于数据量大的情况效果比较明显。

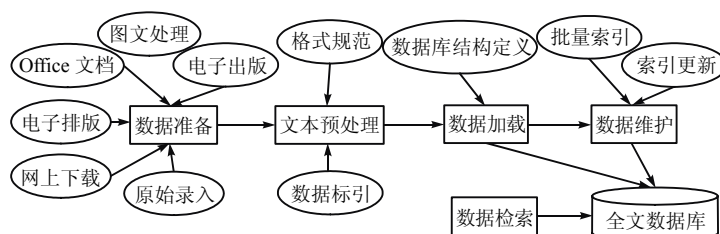


图 7.8 全文数据库的建库流程

2) 文本预处理

(1) 规范格式。当格式多种多样时，应加以整理，使文献的格式规范化。

(2) 数据标引。在建立全文数据库之前，特别是数据加载之前，对非单字索引的正文，利用文字处理软件、专用自动标引软件或人工对数据进行的标引，包括外部属性的著录和内容特征的分类、主题标引等。

3) 数据加载

数据准备好以后，便可以加载到数据库文件中去了。加载数据可采用单篇方式或批量方式。单篇方式一次加载一篇，适于平时文献随时加载的情况；批量方式一次加载多篇，适于集中大量加载的情况。

4) 数据检索

数据库建立起来之后，便可根据全文检索系统提供的检索功能对数据库中的记录进行检索，并返回题录和全文信息。

5) 数据维护

全文数据库建立以后，需要经常对数据库的内容进行索引、更新、追加和清理，以保证数据库的实用性、有效性和时新性。对全文数据库的维护通常包括：全文数据库的结构定义内容，全文数据库的数据内容，全文系统中所用词表、存储空间利用统计及调整。

2. 全文数据库的开发技术

全文数据库开发涉及数据库、文字处理、存储、压缩、检索、知识链接、网络通信、信息安全等技术的支撑，这些技术的发展大大推动了全文数据库的建设。

1) 数据库技术

数据库是一种有效管理和存取大量数据资源的计算机技术。全文数据库一般由一个变长的主文件和一个在索引文件控制下的倒排文件组成。索引文件和倒排文件在物理上是分开的，逻辑上可组合成倒排索引文件。检索时，由索引文件指向倒排文件，倒排文件指向主文件。数据库结构定义是全文数据库建设的数据结构规范，包括记录层次的确定、字段的确定、索引方式的确定等。

2) 信息资源加工技术

全文数据库的信息资源加工技术主要包括信息资源的数字化技术、标引组织技术和知识链接技术。

信息资源数字化尤其是全文数字化技术是全文数据库资源建设的关键，主要由扫描技术、OCR 识别技术、文档生成技术、图文处理技术等组成。

全文数据库的标引技术主要指信息特征的自动抽取和信息内容的自动标引等。信息特征的自动抽取是利用信息抽取技术自动识别信息的标题、摘要、篇章结构、作者、机构、地名等信息并进行标注以提供这方面的特性检索；信息内容的自动标引是对信息内容特征的揭

示,包括自动分类和关键词标引技术,对全文数据库进行适当标引是有益的、必要的。

知识链接是根据知识体之间的关联关系将它们联系起来的过程、方法和技术。ISI 的 Web of Science 引文索引数据库是利用知识链接技术的早期代表。现在,诸如中国知网、万方数据、维普等全文数据库都将知识链接技术应用到其数据库内,包括知识元链接、引证链接、相似文献链接、机构链接、基金链接、作者链接等具体链接方式。

3) 存储与压缩技术

全文数据库的海量化,要求实现海量、高速的存储和高效、无损的压缩。存储介质、存储设备不断更新换代,传统的存储设备,如光盘、磁盘、光盘塔、磁盘阵列容量越来越大。而网络连接存储 NAS、存储区域网络 SAN 和存储网络 ISCSI 等海量网络化存储技术的涌现,则为基于 Web 服务的全文数据库的海量存储提供了更好的解决方案。数据压缩是对数据进行重新编码,以减少所需存储空间。对图像、声音、视频等数据的压缩是多媒体全文数据库得以大规模发展的前提。

4) 计算机检索技术

海量数据的搜索速度和效率是全文数据库检索系统面临的巨大挑战。全文检索、图像搜索、语音检索、智能检索,以及跨库、跨平台和跨语言检索等计算机检索问题都得到逐步解决。

5) 安全技术

信息安全是全文数据库建设中不容忽视的一个问题,包括信息系统的安全、信息内容的安全和用户信息安全等,目前常用的安全措施有权限管理、信息加密等。

7.3.3 全文数据库举要

1. 西文全文数据库举要

1) ProQuest 全文数据库

ProQuest 是美国 ProQuest Information and Learning 公司(原 UMI 公司)经营的全球性全文检索和传递系统,其数据库覆盖了商业、金融、新闻、科技、医学、综合参考及人文社会科学等多个学科领域,包含期刊、报纸、参考书、书目、索引、地图集、绝版图书、档案、学位论文和会议论文等各种类型的信息服务,提供数字化、缩微胶片和印刷版三个类型产品。

(1) 主要资源内容。

ProQuest 系统目前有 70 余个全文数据库,收录内容偏重学术性,全文所占比例较高,适合大学和研究机构使用。其主要产品包括如下几种。

① 商业信息数据库 (ABI/INFORM, Abstracts of Business Information): 世界著名的商业及经济管理期刊论文数据库,收录有财会、银行、商业、计算机、经济、能源、工程、环境、金融、国际贸易、保险、法律、管理、市场、税收、电信等主题的 2600 多种商业期刊,涉及这些行业的市场、企业文化、企业案例分析、公司新闻和分析、国际贸易与投资、经济状况和预测等方面,其中全文刊物有 1800 余种,其余为文摘,有图像。收录时间最长的期刊始于 1986 年。

② 学术研究图书馆 (ARL, Academic Research Library): 综合参考及人文社会科学期刊论文数据库,涉及商业经济、教育、保护服务/公共管理、社会科学及历史、计算机、科学、工程、传播学、法律、军事、文化、医学、卫生健康及其相关学科、生物科学/生命科学、艺术、视觉与表演艺术、心理学、宗教与神学、哲学、社会学及妇女研究等学科,收录 3862 种期刊和报纸,其中全文刊有 2623 种,有图像。可检索 1971 年以来的文摘和 1986 年以来的全文。

③ 应用科学与技术数据库 (ASTP, Applied Science & Technology Plus): 收录应用科学

和工程技术学科的 700 余种期刊, 其中全文刊 300 多种, 主题范围涉及计算机科学、工程技术、物理、电讯、航空航天、交通运输等。收录文献类型有图书评论、新闻、访问、统计资料、图表等 26 种, 每周更新。

④ 博硕士论文数据库 (PQDT, ProQuest Dissertations & Theses): 收录了欧美 2000 余所大学文、理、工、农、医领域的 200 万篇博士、硕士学位论文的摘要及索引, 其中博士论文摘要 350 字左右, 硕士论文摘要 150 字左右, 1997 年以后的博士论文提供前 24 页的全文, 收录起始于 1861 年。PQDT 数据库有三个版本, 分别以 A、B、C 标识来区别不同的学科范围, PQDT-A 为人文社会科学类, PQDD-B 为科学及工程类, PQDD-C 为综合类。PQDT 是世界著名学位论文数据库, 是学术研究中十分重要的参考信息源。

(2) 数据库检索。

ProQuest 检索系统有如下特点。

① 开发历史较早, 发展成熟, 检索功能完善, 具备复杂检索、简单检索、浏览、索引、自然语言检索、二次检索等多种检索功能; 提供十多个检索入口; 可运用布尔逻辑检索、截词检索、位置算符、嵌套运算、限定检索等多项检索技术。

② 提供检索结果的多种处理方式, 可以浏览并标记记录, 以打印、存盘、E-mail 发送三种方式输出。

③ 全文输出完整, 文中的表格、图表、图像与全文一样, 多数文献具备文本和 PDF 文件两种输出格式供选择。

④ 在检索语言上, 可进行自然语言检索, 有主题词表供用户浏览和检索使用。

⑤ 界面友好, “帮助”文件完整, 易学易用。

2) EBSCOhost

EBSCOhost 是一个集文摘型和全文型数据库于一体, 涵盖科技、商业、教育、文学、传媒等领域的综合型检索系统。EBSCOhost 目前有近百个在线文献数据库, 包括近 3000 种期刊全文。

(1) 主要资源内容。

① 学术期刊全文数据库 (ASE, Academic Search Elite): 包括工商经济、资讯科技、人文科学、社会科学、通信传播、教育、艺术、文学、医药、通用科学等领域近 3000 种期刊, 其中全文刊 1500 多种, 最早收录时间为 1990 年, 有图像、图表。

② 学术期刊集成全文数据库 (ASP, Academic Search Premier): ASE 数据库的全文刊增加到 3200 种后, 数据库名称改为 ASP, 共包括 7699 种期刊, 其中提供全文的期刊有 3971 种, 是当今世界最大的多学科学术期刊全文数据库, 专为研究机构设计。

③ 商业资源全文数据库 (BSE, Business Source Elite): 包括 1600 种期刊的索引、文摘, 其中 1000 余种全文刊, SCI 收录的核心全文刊有 238 种; 涉及的主题范围有国际商务、经济学、经济管理、金融、会计、劳动人事、银行等。

④ 商业资源集成全文数据库 (BSP, Business Source Premier): 是 BSE 数据库的升级版, 包括 4000 余种期刊的索引和文摘, 其中全文刊约 3600 种, 全文最早收录时间为 1990 年, 有图像。

(2) 检索系统。

EBSCOhost 系统提供简单检索、高级检索和浏览等检索方式, 在检索功能、检索语言、检索技术、检索结果等方面基本具备与 ProQuest 系统同样的优点; 界面设计友好, 有许多小功能方便用户使用, 例如, 可以链接到用户所在的图书馆主页和一些相关数据库, 可进行不同语种之间的翻译等。

3) 其他外文全文数据库

(1) Elsevier Science 是世界上公认的高品位学术出版公司,也是全球最大的出版商,已有 100 多年的历史。除了出版图书外,还是当今世界最大的学术期刊出版商,出版的期刊大多数都是核心期刊,并且被世界上许多著名的二次文献数据库收录。1997 年,Elsevier 公司建立 Science Direct 全文数据库,通过网络提供服务,内容覆盖数学、物理、化学、医学、生命科学、商业及经济管理、计算机科学、能源科学、环境科学、材料科学、社会科学等众多学科。

从 2000 年起,由 CALIS 工程中心、清华大学牵头组织集团购买该数据库。目前,国内已有 100 多所高校、中科院所及国家图书馆等机构加入,中国用户可以访问的期刊全文自 1995 年起。

(2) SpringerLink 电子期刊数据库由德国施普林格(Springer-Verlag)出版集团出版。Springer 公司以出版学术性出版物而著名,是出版图书、期刊、工具书的综合性出版公司,也是较早将纸本期刊做成电子期刊发行的出版商之一。

荷兰 Kluwer Academic Publisher 是具有国际性声誉的学术出版商。它出版的期刊、图书一向品质较高,备受专家学者的信赖和赞誉。2004 年,Kluwer 并入 Springer, Springer 电子期刊与 Kluwer 电子期刊整合到一个平台上,通过 SpringerLink 发布。目前 SpringerLink 中的期刊种类已达到了 2400 多种,学科覆盖行为科学、生命科学、商业和经济、化学和材料、计算机机科学、地球和环境科学、工程技术、人文及社会科学、数学、医学、物理和天文。

(3) Wiley InterScience 电子期刊。

John Wiley 公司成立于 1807 年,是美国最古老的出版公司之一,也是全球最大的独立出版机构。作为一个综合性出版公司,出版有图书、期刊、各类参考工具书等,其中期刊以质量和学术性见长。目前,John Wiley 上网的电子期刊 700 多种,具体涉及的学科包括:生命科学与医学、数学与统计学、物理、化学化工、地球科学、计算机科学、工程学、商业管理、金融学、教育学、法律、心理学。

John Wiley 电子期刊的检索系统为 InterScience,将 John Wiley 出版的电子图书、期刊和参考工具书放在同一系统平台上,既可分类型检索,又可跨库检索,方便用户。

(4) 威尔逊精选全文数据库(Wilson Select):是 H. W. Wilson 公司的全文数据库,包括著名“威尔逊普通科学文摘”、“人文科学文摘”、“读者指南文摘”和“威尔逊商业文摘”等数据库收录的 1600 种期刊的全文,涉及会计、广告、审计、银行、保险、国际贸易、投资、计算机、经济、工程、市场、行政管理、房地产、金融、税收、通信、交通等领域。

(5) LEXIS-NEXIS 学术大全数据库(Academic Universe):收录有法律信息、案例、新闻、商业金融信息、政府规章制度、各类参考资料等,包含期刊、报告、政府出版物、新闻快讯等 6000 余种出版物,其中约 90%有全文或部分全文。其中法律法规方面的数据库是该数据库的特色信息源,在法律界具有很高的知名度和很大的影响力。

2. 中文全文数据库举要

1) 中国知网

CNKI(CNKI, China National Knowledge Internet)是由清华同方光盘股份有限公司、中国学术期刊(光盘版)电子杂志社和清华大学光盘国家工程研究中心共同启动和实施的一项国家信息资源建设工程。CNKI 整合了目前我国现有的期刊、博硕士学位论文、会议论文、报纸、年鉴、工具书、专利、科技成果、国内外标准、哈佛商业评论、国学宝典、外文(NSTL)、法律库等多种类型的文献资源,是国内最为成熟的数字资源数据群,全球最大的中文知识门户。

(1) 主要资源内容。

① 中国学术期刊网络出版总库, 收录国内 8000 多种重要期刊, 以学术、技术、政策指导、高等科普及教育类为主, 同时收录部分基础教育、大众科普、大众文化和文艺作品类刊物, 内容覆盖自然科学、工程技术、农业、哲学、医学、人文社会科学等各个领域。截至 2013 年 3 月, 全文文献总量 3690 万篇。产品分为十大专辑: 理工 A、理工 B、理工 C、农业、医药卫生、文史哲、政治军事与法律、教育与社会科学综合、电子技术与信息科学、经济与管理等。十专辑下分 168 个专题和近 3600 个子栏目。收录全文主要始于 1994 年, 部分刊物甚至回溯至创刊文献。

② 中国硕士、博士学位论文全文库, 是目前国内相关资源最完备、高质量、连续动态更新的中国硕士、博士学位论文全文数据库。截至 2013 年 4 月, 累积 1984 年以来博士学位论文全文近 20 万篇, 硕士学位论文全文数据库 170 万篇。

③ 中国重要报纸全文库, 收录 2000 年以来国内公开发行的 500 多种重要报纸刊载的学术性、资料性文献。至 2012 年 10 月, 已累积报纸全文文献 1000 多万篇。

④ 中国重要会议论文全文刊, 收录我国 2000 年以来国家二级以上学会、协会、高等院校、科研院所、学术机构等单位的论文集, 年更新约 10 万篇论文。至 2008 年 5 月, 累积会议论文全文文献近 90 万篇。

其他诸如《中国统计年鉴数据库》、《中国专利全文数据库》、《中国标准数据库》、《中国图书全文数据库》等不再赘述。CNKI 中心网站及数据库交换服务中心每日更新数据, 各镜像站点通过互联网或卫星传送数据也实现每日更新, 专辑光盘每月更新, 保持数据的更新和增量发展。

(2) 检索系统。

① CNKI 数据库提供了强大的检索功能, 既支持单库检索也提供跨库检索; 检索入口丰富多样, 不同的数据库根据资源内容特色设置了不同的检索点, 如期刊论文的检索点有刊名、标题、作者、作者单位、关键词、文摘、引文、分类及基金等, 而会议论文的检索点则设为会议名称、会议录名称、会议地点、主办单位、标题、作者、作者单位、关键词、文摘、引文、分类等。

② 统一导航。提供了以 168 学科导航为基础的学术文献导航, 通过使用文献导航, 可控制检索的学科范围、提高检索准确率及检索速度。168 学科导航既可作为检索范围控制, 也可直接查看/浏览每个导航类目下的文献。

③ 不同检索方式。根据信息检索的需求, 提供了快速检索、标准检索、高级检索、专业检索、引文检索、作者发文检索、科研基金检索、句子检索、工具书及知识元搜索、文献来源检索等面向不同需要的检索方式。

④ 检索结果分组。检索结果页面将通过检索平台检索得到的检索结果以列表形式展示出来, 并提供对检索结果进行分组分析, 检索结果分组类型包括: 学科类别、期刊名称、研究资助基金、研究层次、文献作者、作者单位、中文关键词。

⑤ 检索结果排序。除了分组筛选, 还为检索结果提供了发表时间、相关度、被引频次、下载频次、浏览频次等排序方式。

⑥ 知识节。CNKI 提供单篇文献的详细信息和扩展信息浏览的页面被称为“知网节”。它不仅包含了单篇文献的详细信息, 而且还是各种扩展信息的入口汇集点。这些扩展信息通过概念相关、事实相关等方法提示知识之间的关联关系, 达到知识扩展的目的, 有助于新知识的学习和发现, 帮助实现知识获取和知识发现。

⑦ 在线检索词库, CNKI 提供刊名、作者、作者机构、关键词等索引词库, 可在线浏览选择所需用的检索词, 极大地方便了用户的各种需求。

2) 万方数据资源系统

万方数据资源系统是一个以科技信息为主,集经济、金融、社会、文化、教育、卫生等各行业信息于一体,以 Internet 为网络平台的现代化、网络化大型科技、商务信息服务系统。它包括 110 余个数据库,归属 9 个类别,内容涉及自然科学和社会科学各个专业领域。收录范围包括期刊、会议、文献、题录、报告、论文、标准专利、连续出版物、工具书、最新科技成果等,主要产品有:中国学位论文数据库、数字化期刊群、中国学术会议论文数据库、中国标准全文数据库、中国法律法规全文库、中国专利全文数据库、科技信息子系统和商务信息子系统等。

3) 其他中文全文数据库

(1) 中国科技期刊数据库。

重庆维普资讯公司推出的《中文科技期刊数据库》是一个功能强大的中文科技期刊全文检索系统,是国内最大的综合性文献数据库之一,目前它收录了中国境内历年出版的中文期刊 12000 种,全文 3000 余万篇,分为三个版本(全文版、文摘版、引文版)和 8 个专辑(社会科学、自然科学、工程技术、农业科学、医药卫生、经济管理、教育科学、图书情报)定期出版发行,2000 年以后每年出版文献 90 万~100 万篇,提供一般检索、传统检索、高级检索、分类检索、期刊导航等多种检索功能。

(2) 超星数字图书馆。

超星数字图书馆是国家“863”计划中国数字图书馆示范工程项目,由北京世纪超星信息技术发展有限责任公司投资兴建,收集了国内各公共图书馆和大学图书馆以超星 PDG 技术制作的数字图书,是目前中国最大的网上数字图书馆,1998 年开始提供网上检索服务。超星数字图书馆包含图书资源近百万种,涵盖《中图法》22 大类,包括文学、历史、法律、军事、经济、科学、医药、工程、建筑、交通、计算机、环保等。

7.4 异构数据库整合与导航

随着一个个不同类型、不同学科、不同平台的信息资源系统的建立,由于各信息资源系统之间相对独立、相互封闭,形成了一个个信息孤岛,严重阻碍了人们对信息的获取。我国已经形成了世界上最大规模的信息孤岛,迫切需要消除信息孤岛、整合数字资源,实现信息资源共享。

数字资源整合,是指依据一定的需要和要求,通过中间技术,把不同来源、不同通信协议的信息完全融合,使不同类型、不同格式的数字资源之间实现无缝链接。通过整合的数字资源系统,具有集成检索功能,是一种跨平台、跨数据库、跨内容的新型数字资源体系。

异构数据库资源整合和资源导航整合是信息资源整合最主要也是最常用的两种方式。

7.4.1 异构数据库整合

1. 数据库现状:问题的提出

目前,许多信息机构大量购置或自建数字化资源,包括各种电子期刊数据库、电子图书数据库、自建特色数据库等,这些形式各异的数据库在给使用者提供丰富信息的同时也带来了信息检索的不便。用户若要查询或获取某一信息,往往需要依次进入各个电子资源的检索界面进行检索,并且要对各个数据库的检索规则有足够的认识,方可获得所需信息。随着数据库资源的不断增加,这些因开发语言和操作系统平台不同、数据库管理系统平台不同、数据模式和数据语义不同而导致的数据库异构问题日益显著。实现信息系统互操作、整合不同

数据库资源是实现信息资源整合的一个主要方面。

异构数据库整合实际上是建立统一检索平台，并利用超文本技术，在不同的信息资源之间进行链接，将原本相互独立但互为联系的信息资源整合在一起，使之形成一个互动的有机整体，用户只须通过同一界面，即可迅速查到并获取自己所需信息。

异构数据库整合平台在图书馆网络服务平台和一些大型数字信息资源服务系统中已广有应用。商业化的产品有 MetaLib、WebFeat、MetaSeach Solution 等异构集成检索系统，以及图书馆等研究机构开发的 FlashPoint、NLM Gateway 和 SiteSearch 等。国内的北京大学图书馆、清华大学图书馆、上海交通大学图书馆、中国科学院文献情报中心等一批技术实力雄厚的图书馆已开发出一些异构数据库统一检索平台。如北京大学图书馆把目前拥有的 200 多个数据库、近 2 万种电子期刊和 10 余万种电子图书整合在一个统一检索平台上，提供异构系统的跨库检索服务，用户可按学科、按数据库名称、按文种同时检索多个数据库的资源（图 7.9）。CNKI 也实现了对异构数据库资源的整合，其知识网络服务平台 KNS 就是一个资源整合服务系统，它提供统一的跨库检索平台，允许同时在《中国期刊全文数据库》、《中国优秀博硕士学位论文全文数据库》、《中国重要会议论文全文数据库》、《中国重要报纸全文数据库》、《中国图书全文数据库》、《中国引文数据库》等 CNKI 源数据库中同时检索（图 7.10）；同时，通过数据库间的引文链接，实现了《中国期刊全文数据库》、《中国优秀博硕士学位论文全文数据库》、《中国重要会议论文全文数据库》、《中国重要报纸全文数据库》等之间的互引用链接，即期刊库论文如果引用了学位论文库的文章，期刊库检索结果中会将学位论文库的该篇文章作为参考文献提供链接，并可点击链接查看文章题录，进而获取原文，反之亦然，构建了异构数据库间的知识网络，实现了真正意义上的异构数据库整合。



图 7.9 北京大学图书馆信息资源统一检索平台

在信息资源日益丰富的信息环境下，实现异构数据库的整合具有重要意义。

(1) 有助于解决资源利用率与服务效率低下的问题。将数字资源按照统一的标准进行组织、整序与整合，能发挥各种资源的最佳使用效益，形成一个一体化的知识服务网络，发挥服务的整体效益。

(2) 有助于解决用户使用与用户培训的困难。异构数字资源整合之后，把各种信息资源透明无缝地连在一起，让用户感觉如同只在一个资源系统中操作，学习和使用起来较为方便。

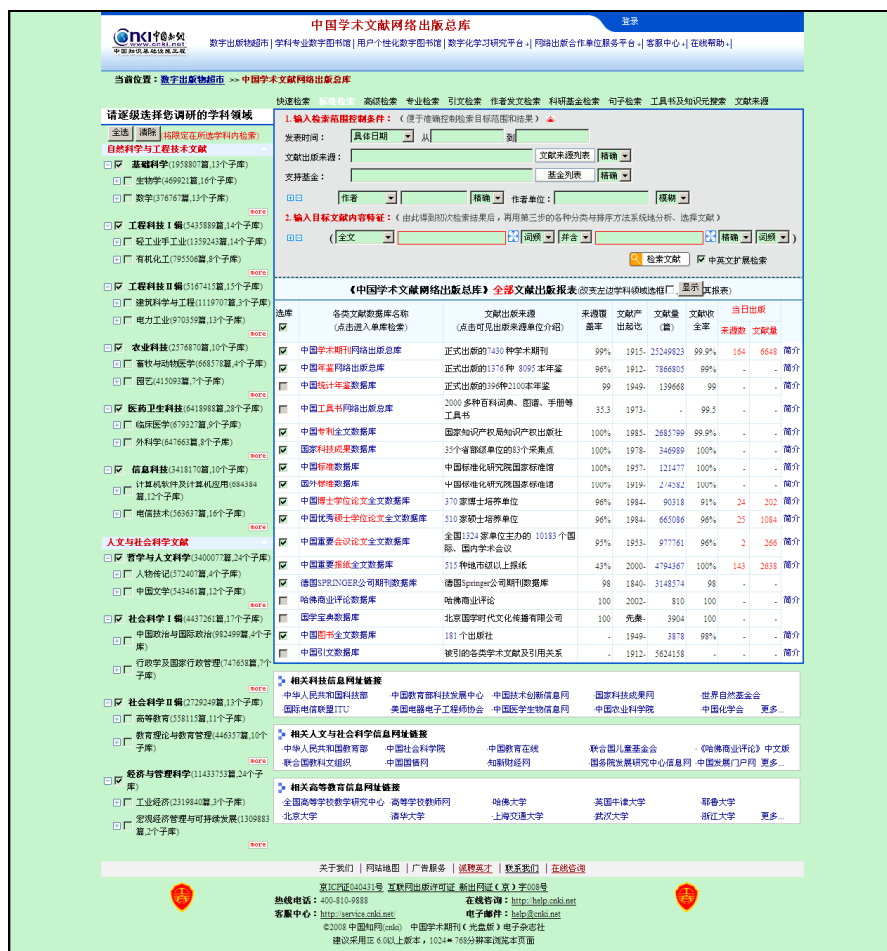


图 7.10 CNKI 异构数据库跨库检索平台

(3) 解决资源管理的困难。不同的系统、不同的平台、不同的配置、不同的存储设备、不同的标准等给系统管理人员带来了很大的麻烦。通过资源的整合，管理人员只需建立一个统一的检索平台就能把各种数据库揭示给用户。

2. 实现整合的协议和标准

要整合异构信息资源，实现“一站式”服务，需要一起遵循相应的协议或标准。当前，常用的异构数据库资源整合协议标准有 Z39.50、OAI、OpenURL、LDAP 等。

1) Z39.50 协议

Z39.50 协议是信息检索应用服务定义和协议规范 (Information Retrieval Application Service Definition and Protocol Specification) 的简称，最初由美国国会图书馆等机构为解决书目信息检索系统之间的通信问题而开发。1988 年正式推出第一版，并成为美国国家标准，截至 2003 年已发布了四版；国际标准化组织于 1996 年将其采纳为国际标准，定名为 ISO 23950。

Z39.50 的目的是为了信息系统的开放互连，由于各信息系统分别采用各自的数据库软件，数据描述格式、访问方式等都各不相同，必须为各自数据库系统建立一个抽象、通用的用户视图，将各个系统的具体实现映射到抽象模型上，才能使不同的系统在一个相互理解的、

标准的通信平台上进行交互。Z39.50 采用了客户机/服务器的灵活架构,通过源端和目标端实现信息交换。Z39.50 主要的服务机制包括 11 项内容:初始化、查询、获取、删除结果集、访问控制、账户与资源管理、排序、浏览、解释、扩展服务、终止等。Z39.50 的工作流程是:客户端作为查询请求的发动者发出检索请求,随后服务器端接收并分析检索式,接着从系统数据库内搜寻出满足检索条件的记录,最后将检索结果返回客户端。图 7.11 是一个典型的基于 Z39.50 协议的异构数据库统一检索平台结构。

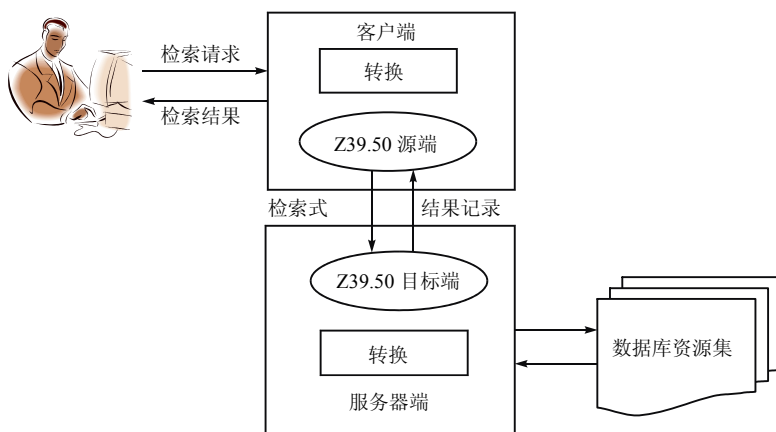


图 7.11 基于 Z39.50 协议的异构数据库统一检索平台结构

Z39.50 实现了信息查询和提取过程的标准化,通过 Z39.50 协议,用户可以通过网络对不同计算机上的信息进行检索,而不必关心这些信息是如何存储和组织。但受当时技术环境的影响,它采用的是一种比较封闭的方式,是一个结构相对严谨而又复杂的数据库通信接口协议,具有一定的技术复杂度,且与 HTTP 不兼容。因此,业界人士开始想方设法改造 Z39.50 协议,在 2001 年推出了适用于 Web 资源的 ZING (Z39.50 International:Next Generation),称为“下一代 Z39.50 协议”,包括 SRW/SRU、CQL、ZOOM、ez3959 和 ZeeRex 五个部分。ZING 一方面可以看成是 Z39.50 各种功能在新的网络协议和应用模式下的发展,另一方面是一种简化,淘汰了 Z39.50 中没用和无意义的内容。

2) OAI 协议

OAI 协议 (OAI-PMH, Open Archives Initiative Protocol for Metadata Harvesting) 是一种独立于应用的、能够提高 Web 上资源共享范围和能力的互操作协议标准。OAI 协议最大的特色是通过相对简单的、独立于应用程序外的元数据收割协议,来实现异构、分布元数据资源之间便捷的互操作,即基于元数据收割实现异构信息资源的整合。

OAI 包含两类角色:数据提供方 (DP) 和服务提供方 (SP)。DP 负责元数据的生成和发布,使之符合 OAI 协议;SP 负责元数据收割和提供统一检索等增值服务。OAI 的核心思想非常简单:在 OAI 协议的基础上,SP 使用 OAI 规定的 6 个命令动词——Identify、ListSets、ListMetadataFormats、ListIdentifiers、GetRecord、ListRecords,将不同 DP 的异构数据库的元数据收割 (Harvest) 至本地,SP 集中精力对收割来的元数据进行进一步的分析整理,开发高效统一的检索平台,在本地服务器上检索,使复杂问题简化,从而提高了检索质量、减少了用户等待时间,OAI 基本架构如图 7.12 所示。因此,在跨库检索中,OAI 协议不失为一种低成本、简便灵活的手段。

目前,国内外主流跨库检索系统都已经支持 OAI 协议。如国外的 MetaLib、Encompass、Webfeat 等知名的跨库检索系统;国内中国科学数字图书馆 (CSDL) 开发的跨库集成检索系

统也提供了 OAI 连接器, 中国高等教育文献保障系统 (CALLS) 的统一检索系统也支持 OAI 协议。由此可见, OAI 协议将在跨库检索中得到普遍应用。

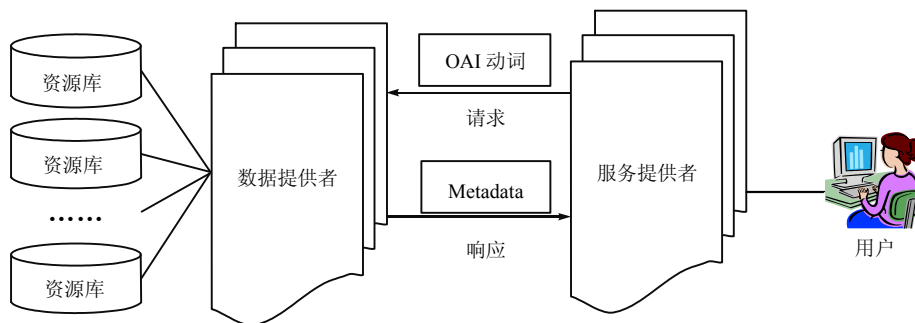


图 7.12 OAI 基本架构

3) OpenURL 协议

OpenURL (Open Uniform Resource Locators, 开放的统一资源定位器) 是于 1999 年由比利时 Ghent 大学的 H. 萨姆堡尔 (Herbert Van de Sompel) 及其同事在研制 SFX 系统时提出的, 目的是把不同来源和不同通信协议的信息源及相关服务融合在一起, 实现不同类型、不同格式和异地分布信息资源的无缝链接。这是一种开放的信息资源与查询服务之间的通信协议标准, 是开放的、上下文相关的链接框架, 它提供了一套在信息服务者之间传递对象元数据的公共语法, 允许信息源公开自己的链接接口, 实现链接信息源和链接服务器之间的信息传输, 从而实现异构数据库之间的互操作, OpenURL 基本运作方式如图 7.13 所示。

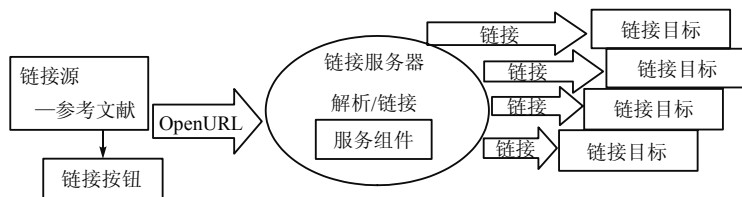


图 7.13 OpenURL 基本运作方式

OpenURL 协议通过链接将用户指引到合适的资源。链接服务器定义了用户的上下文环境, 当链接服务器接收到一个输入的 OpenURL 后, 就会根据该用户所在机构的馆藏向其提供服务。OpenURL 标准允许已经检索到文献线索的用户即时获取到该文献资源, 如从文摘库链接到全文库, 实现原文的即时获取。目前基于 OpenURL 框架的参考链接系统产品多达几十种, 其中较成熟、应用较广的系统有 SFX、WebBridge、LinkFinderPlus。

这些协议从系统互连、数据整合和信息获取三个方面解决了分布、异构数据资源的检索、整合和共享问题, 它们既可以单独使用又可以与其他协议一起协同工作, 提供异构数据库的整合服务。

3. 异构数据库整合的技术

对于异构数据库的整合目前主要从两个方面入手: 一是从信息源上入手, 要求源数据库的输入/输出遵循相应的通信协议, 基于共同协议来实现跨库检索; 二是承认信息源的差异, 开发一种能够适应不同检索模式的检索代理程序来实现异构数据库的整合, 是一种基于中间件的跨库检索。

关于异构数据整合的协议在上文已有论述，此处不再论述。构建中间件是异构数据库整合的常用技术，当用户提出检索请求后，其请求被交给服务器端的一个 Agent 程序。Agent 针对不同的数据库，将用户请求转化为符合其规定的格式，然后将请求发送到各数据库；在得到数据库的返回结果后，Agent 再将不同数据库的结果转化为统一的格式，并发送到浏览器端显示给用户，与元搜索引擎的机理颇为相似。清华同方的异构数据库统一检索平台就是比较成熟的中间件模式的异构数据库整合模式。

异构数据库整合涉及的技术较多，在当前技术和主流平台中，XML 是广泛应用的数据交换语言接口；数据仓库在信息的组织和信息提供方面具有强大的功能；.NET 技术在分布式应用中表现出优秀的特性；新兴的 SFX 技术将数据库的元数据定义为标准的 OpenURL 格式，提供不同数据库资源的互操作功能，使各类复杂的数据和信息之间的关联变成简单的链接，它不仅能完成从二次文献到全文的链接，还能实现从文摘到文摘、全文到全文的链接，使所有的异构资源完全融合为一个整体；Web Services 是由 W3C 组织定义的一个采用 XML、通过 URL 来发布接口和应用绑定的软件体系结构，它采取简单的、易于理解的标准 Web 协议作为组件界面描述和系统描述规范，完全屏蔽了不同软件平台的差异，能够统一地封装信息、行为、数据表现及商务流程，是异构平台集成的最佳手段，在异构资源整合方面具有卓越的性能。图 7.14 展示了一个异构信息系统整合的技术框架，它分别从资源层和应用层来解决异构数据的整合。目前异构数据库的整合主要从应用层入手，基于软件或基于协议实现整合；而基于元数据和知识组织系统的资源语义层面的整合随着关联数据的发展开始起步。

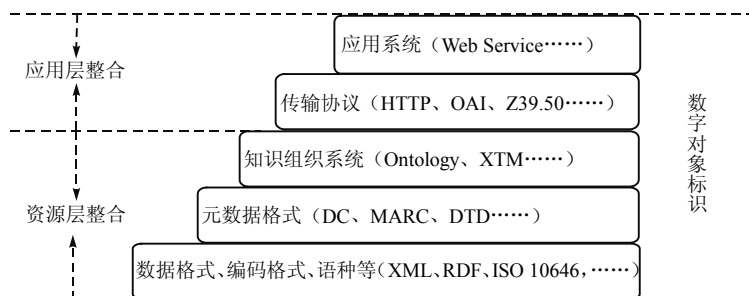


图 7.14 异构信息系统整合的技术框架

7.4.2 数据库导航

1. 数据库导航的概念及意义

随着购进和自建数据库资源的日益增多，如果仅仅把数据库按名称罗列于网页上，不能很好地揭示数据库内容，造成用户使用的随机性和盲目性，数据库利用率不高。因此，越来越多的信息机构开始重视数据库的导航工作。数据库导航是另一种信息资源整合方式，即为信息机构拥有的电子资源数据库建立导航系统，实现对众多数据库资源合理有效的排序和整合，使资源能够清晰、有序地供用户选择检索利用。

建立数据库导航系统实际上是通过数据库的剖析和分离，根据学科或主题将其归类重组，覆盖多学科数据库重复归入多个类目，各学科或主题类目下再按字顺排列，既有按数据库名称字顺和学科浏览功能，又有按数据库名称关键词或学科/主题检索功能，每一种数据库都有超文本链接点，指引进入目标数据库。这样，用户进入数据库导航系统，单击相关学

科或主题,便可展现该学科/主题相关的所有数据库,包括二次文献和全文文献数据库,杜绝漏检的情况。建立数据库导航系统是对信息资源的一次重组和导引。

2. 数据库导航的建设过程

数据库导航系统的构建方法与其他信息导航系统构建方法相似,主要包括以下几步。

1) 数据库资源的收集

在建立导航系统之前,先要对所拥有的数据库资源进行收集整理,包括购进数据库和自建数据库资源,以便确定范围。

2) 数据库资源的描述

为了便于用户了解其所需的数据库资源,必须对数据库所包含的有用信息进行细分和揭示,如数据库名称、国别、语种、收录时间范围、收录内容范畴、数据库网址、资源类型、检索方法和步骤等。

3) 数据库资源的分类

最主要的数据库导航方式是按学科导航,因此,需要根据数据库资源所涉及的学科范围和用户需求的主题来确定导航系统中的学科分类体系,并据此对数据库资源进行分类标引和主题标引。

4) 数据库资源的组织序化

通过对数据库信息的揭示,生成数据库资源的元数据,然后将数据库资源按照某个或多个描述字段进行序化组织,如按照数据库名称的字顺排列,按照资源类型排列、按照分类的等级结构排列等,在导航首页上进行展示。

5) 导航检索系统的建立

构建一个功能齐全的检索平台是数据库导航成功的关键,检索途径是否完备关系到站点的检索能力和检索便利性,不论采用什么系统都应着眼于提供多元化的检索功能。可以把数据库资源名称、资源类型、语种、主题、揭示层次、数据更新日期、推荐人、责任人、资源说明等作为检索点,让用户在检索时有尽可能多的选择。

6) 导航系统的更新维护

随着数据库资源情况的变动,必须及时更新维护数据库导航系统。更新维护是保障数据库导航系统质量的后续工作,是保障数据库有较高利用率的重要手段。这项工作主要从两个方面展开:一是数据源的更新,及时更新新增数据库资源,删除不再提供使用的数据库链接;二是通过对用户使用导航访问数据库的情况的跟踪分析和用户评价调研,了解数据库资源的利用情况,为今后导航资源的维护与更新提供取舍依据,并根据用户使用反馈定期对导航页面进行维护。

3. 数据库导航系统举要

目前,大多数馆藏电子资源丰富的信息机构都在其服务站点上建立了本单位数据库导航系统,但由于技术水平和服务水平不一,导航系统的质量也良莠不齐。本书仅列举几个数据库导航系统建设较好的机构,以供参考。

1) 大学图书馆数据库导航系统

近年,随着经费投入增加,各大学图书馆纷纷购进与本校学科发展相关的文献数据库,有的甚至自建特色数据库,这些数据库与图书馆的书目数据库等构成了大学图书馆丰富的数据库资源,对这些数据库的管理与导引也成为图书馆网站建设的主要工作之一。

武汉大学图书馆对其订购的200余种中外网络数据库和光盘数据库及其子数据库建立了功能完善的导航系统,如图7.15所示。通过人工对各个数据库的资源类型、内容、揭示层次、学科类别、名称、出版商、链接地址等进行标引。导航平台则提供按字母、学科类别、文献类型三种方式进行浏览和检索的途径,其中学科类别涉及综合型大学的28个学科门类,文

献类型包括专著、期刊论文、学位论文、报纸、标准、专利、报告等多种文献形式。检索结果的显示包括数据库名、收录年限、收录文献类型、数据库类型、简介、上层数据库、出版商等，并提供链接直接进入相应的数据库。

北京大学图书馆拥有数据库资源 200 余种，其中外文数据库 140 种，中文数据库近百种，在其图书馆网站上提供了一个数据库资源导航系统，可以分别通过学科、语种和字顺三种方式浏览，并支持这三种方式的组合查询，如图 7.16 所示。该导航系统的学科分类设置较细，有些学科深入到二级类目，但也存在一点问题，虽然设计了较细的学科，但该学科下面常常是没有数据库资源的。该导航系统主要对数据库的名称、语种、网址、简介、学科、检索方法、是否提供全文进行了描述，结果展现方式较为简单。



图 7.15 武汉大学图书馆数据库导航系统



图 7.16 北京大学图书馆数据库导航系统

清华大学图书馆数据库导航系统提供了字顺、学科分类、文献类型、语种 4 种方式浏览数据库，分别从数据库名称、年限范围、文献类型、提供机构访问方式、学科分类、简介等方面来描述数据库资源。

从这些大学图书馆数据库导航系统的建设来看，导航的数据库多为本馆所拥有的电子资源数据库，而本馆的藏书书目数据库并未加入导航中；导航的方式以字顺和学科分类浏览为主，部分数据库名称或文献类型提供检索功能；对数据库信息均进行了详略不等的描述。

2) 文献信息服务机构数据库导航系统

除了大学图书馆，一些公共文献信息服务机构数据库的拥有量也比较大，一般也会提供相应的电子资源数据库导航服务。如上海图书馆数据库导航系统（见图 7.17），分别从文献类型、数据类型、学科主题和名称字顺 4 种途径来导引其所能提供的数据库资源，并对数据库名称、载体类型、数据类型和收录时间范围、主题、更新周期、简介、访问次数等信息进行揭示，提供数据库名称、生产厂商名称和数据库所有描述信息的检索。但这类信息机构数据库导航系统仍是针对电子资源数据库设计的，与馆藏纸本图书或纸本期刊的书目数据库仍是分离的。



图 7.17 上海图书馆数据库导航系统

3) 商业信息服务机构的数据库导航系统

一些大型的数据库服务商，如 Dialog、EBSCO、CNKI 等，在其服务网站页面上也会针对自己生产提供的数据库进行导航，有的导航比较简单，只是在页面上按字顺或文献类型简单罗列；有的则很精细，编制面向用户检索需求的分类导航系统，引导用户根据检索主题进入一个或多个数据库中，如 Dialog 系统。Dialog 是全球最大的联机数据库服务系统，包含 600 多种数据库，有全文型、书目型、数值型、指南型等不同类型，涉及综合学科、自然科学、应用科学和工艺学、社会科学和人文学、时事报道和商业经济等多个领域，它的数据库导航系统按照学科或主题进行类别层层划分，到相应的类别或检索入口下会提供一个或多个数据库作为该类信息的检索源，如图 7.18 所示。当用户从“医药”大类进入到“医学研究”这个主题后，导航系统会给出 BIOSIS、CAB、Ei Compendex、EMBASE、INSPEC 等 10 个数据库作为信息源，以这种与用户检索需求相关的方式来推荐数据库。

从目前的数据库导航系统来看，大多是基于数据库层面的一次面向学科或面向文献类型的归类，未能深入到数据库内部来进行内容导航。现在已经有一些机构开始对所收藏的多个数据库内容层面进行解析，深化导航的层次。如图书馆定购的各种电子资源中，增幅最大、

利用率最高的是电子期刊，而这些电子期刊是分散在不同的数据库系统中的，如果将不同数据库中的电子期刊集中起来，建立电子期刊导航系统，将极大地方便用户从整体上了解和利用电子期刊资源，是一种深层次的数据资源导航方式。清华大学、武汉大学等图书馆网页上已经提供了这样的中文、西文电子期刊导航系统，是对电子资源数据库的一次解析与重组。

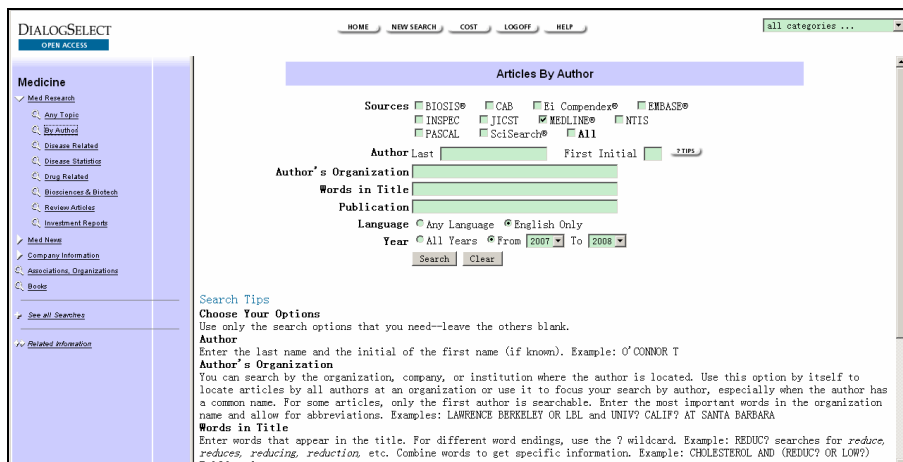


图 7.18 Dialog 的数据库导航系统

7.4.3 学科信息门户

1. 学科信息门户概述

学科信息门户（SIG，Subject Information Gateway），也称主题网关（SP，Subject Portal）、信息网关（IG，Information Gateway）、基于主题的信息网关（SBIG，Subject-Based Information Gateway），它针对特定学科或主题领域，按照一定的资源选择和评价标准及规范的资源描述和组织体系，对具有一定学术价值的网络资源（如网站、文献、工具、服务等）进行搜集、描述和组织，并提供浏览、检索、导航等基本功能。此外，学科信息门户的扩展功能还包括提供一些自创资源（如内部全文文献、专家访谈）、信息聚合 RSS、学术博客、学术论坛等增值服务及个性化服务。总而言之，学科信息门户是一个信息集成整合机制，旨在为用户提供一个方便的学术信息和各种服务的检索和交互学习入口。

国外的学科信息门户建设始于 20 世纪 90 年代，国内稍晚一些，始于 2000 年。国内外的门户系统经过一段时间的建设，已取得一些成就。国外门户系统依靠高校的信息资源和人力技术资源，产生了较大的社会影响，吸引了一些社会资金投资和众多的自愿者参与资源建设，国外著名高校图书馆如耶鲁和普林斯顿，都将 SOSIG、LII 等知名门户的资源纳入各自的馆藏数字资源体系。

国外的学科门户系统往往由两个或多个不同单位共同建设和维护，如 LII 是由图书馆服务与技术联合会、国家图书馆、华盛顿州立图书馆、加利福尼亚数字图书馆、加州大学图书馆等研究机构共同组建的；Intute 由英国曼彻斯特大学、牛津大学、曼彻斯特城市大学、诺丁汉大学、赫瑞瓦特大学、布里斯托尔大学和伯明翰大学 7 所大学共同建立，Intute 设有一个制定战略方针和统筹管理的执行委员会和 4 个为教工和学生提供学科服务的学科小组。我国学科信息门户承建单位比较单一，如 CSDL 由中科院各个院所图书馆承建，CALIS 由高校图书馆系统承建。

学科信息门户致力于将特定学科领域的信息资源、工具与服务集成到一个整体中，为用户提供方便的信息检索和服务入口。从本质上讲，学科信息门户是含有不同分类主题的网页及相关链接的网上图书馆。主要特点是：有较多的人工参与，通过质量标准规范资源的选择，并提供对资源的丰富描述；提供依据学科体系结构和资源类型分类的浏览和检索入口；有对资源的管理和长期发展的政策、元数据应用与标引规范、资源共享与互操作机制等。

学科信息门户与搜索引擎最大的区别在于它提供的信息是高质量的、系统的、受控的。学科信息门户有计划、有针对性地由专业人员搜集信息，对信息源进行质量鉴别后进行取舍，对纳入学科门户的信息使用受控语言和关键词进行必要的内容描述，最后进行系统组织，一般都从学科属性、资源类型角度多重揭示，这些都是搜索引擎无法做到的。通过学术信息门户可以迅速把握一个学科领域的主要信息源，不必用搜索引擎盲目查找，也不会受无用信息的干扰，对科学研究有着重要的作用。

2. 学科信息门户构建

1) 学科信息门户体系结构

从体系架构来看，学科信息门户一般由用户界面层、个性化可定制用户构件层、信息与应用集成层、信息结构层和信息资源层组成，具体层次结构如图 7.19 所示。

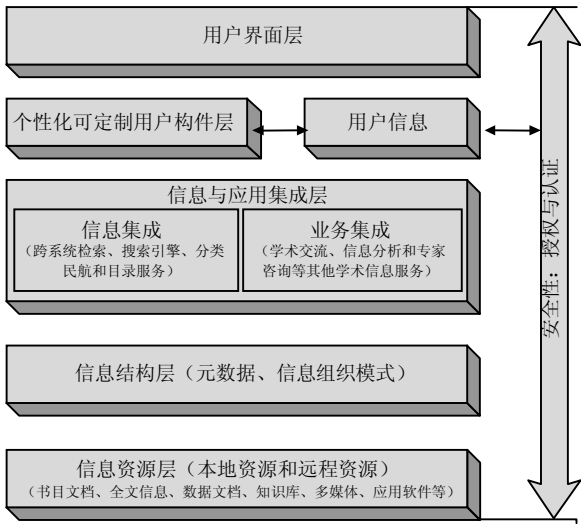


图 7.19 学科信息门户层次结构

- (1) 用户界面层是学科信息门户呈现给用户的表现形式，通常包括跨库检索、搜索框、分类导航、学术交流、专家咨询、个性化定制服务等功能组件。
- (2) 个性化可定制用户构件层主要提供可根据用户信息库中用户的信息需求和权限，调用不同的门户构件，针对不同用户生成不同的页面信息内容。
- (3) 信息与应用集成层包括信息集成和业务集成两块，其中信息集成方面主要涉及跨库检索技术和策略、元搜索引擎和分类导航，而业务集成主要将信息机构的信息分析、专家咨询、学术交流等个性化服务应用集成于此。
- (4) 信息结构层即学科信息门户资源的元数据格式和信息组织方式。
- (5) 信息资源层包含了集成于学科信息门户的各种类型的本地资源和远程资源。

2) 学科信息门户资源建设

资源建设是学科信息门户构建的核心任务，门户资源建设主要包括资源采集、资源描述、

资源组织、资源展示与服务三部分内容。

(1) 学科信息门户资源采集。严格的资源选择是学科信息门户知识性与可靠性的保障,因此资源采集要有明确的目标和策略。在“以用户为中心”的新服务环境下,信息采集时还应充分考虑用户信息需求,通过分析用户需求和偏好来创建内容清单,明确资源的类型、来源和选择标准。目前学科信息门户的资源采集包括人工采集和自动采集两种方式。人工采集方式由人工筛选和专家推荐,能保证信息资源质量,但工作效率低、采集信息不全面,并带有一定的随机性和随意性。自动化采集效率高,采集信息较全面但所收集的信息结果数量大、质量参差不齐。因此,学科信息门户建设中应将自动采集与人工采集相结合,将自动采集的网页进一步进行人工筛选以提高信息采集质量。

(2) 学科信息门户资源描述。网络环境下资源描述多以都柏林元数据为主,这种元数据形式适用于各种网络资源的描述,能有效促进网络资源的组织与获取。在构建学科信息门户时往往需要根据收录资源类型和提供检索方式、服务方式类型来制定更详细的元数据标准体系。

(3) 学科信息门户资源组织。学科信息门户对网络信息的组织,具有搜索引擎等其他网络信息组织模式不可比拟的优势,其原因在于它充分利用规范的知识组织体系来揭示和组织网络信息,弥补了现有模式对专业网络信息资源组织的不足。分类方式和主题方式是学科信息门户建设中最常用的资源组织方式,此外,字顺、时间、地理位置、文献类型等也会作为信息化组织的依据。

(4) 学科信息门户资源展示与服务。学科信息门户应尽可能地提供各种导航工具和检索方式,将网络资源展示给用户,为用户提供信息服务。同时,从网站用户体验考虑,学科信息门户还要考虑网站交互系统的设计,通过即时聊天工具、在线咨询、QA 系统、个人空间、主题论坛等方式来实现网站与用户的交互沟通。而社会化网络的发展,各种 Web2.0 技术和服务也应融入到学科信息门户建设中,以博客、微博等方式建立用户、网站建设者、服务人员的交互,以各种社会化分享方式来实现资源推荐和评价。

3. 学科信息门户举要

1) 中国科学院国家科学数字图书馆

CSDL (CSDL, Chinese National Science Digital Library) 是服务于中国科学院全院网络的科技信息服务环境,通过 CSDL,可以登录互联网免费使用 30 多个科学文献数据库,并获得随易通、文献传递、参考咨询和跨库检索等近 10 项网络化服务。CSDL 收录的资源类型包括:组织机构、数据库、文献、专利、新闻组、期刊、讨论组/邮件列表、图书、软件、科技基金、名录、会议、参考工具、公司;CSDL 提供 13 种外文全文数据库,覆盖了 2863 种核心期刊、6409 种西文会议录;内容涉及数学、物理、化学、生命科学、社会科学、天文学、电气与电子学、计算机科学等领域。此外,CSDL 还提供了由其相关院所自建的化学学科信息门户、生命科学信息门户、数理学科信息门户、图书情报信息门户、资源环境信息门户等专题性信息门户。

2) CALIS 重点学科网络资源导航门户

CALIS 重点学科网络资源导航门户由中国高等教育文献保障系统 (CALIS, China Academic Library & Information System) 的成员共同建成,现已建成 265 个学科导航系统,其学科几乎覆盖了社会科学 (75 个)、自然科学 (190 个) 的各个学科领域。其中工程技术类重点学科导航由上海交通大学图书馆负责、文理类由北京大学图书馆负责、农业类由中国农业大学图书馆负责、医学类由上海医科大学图书馆负责。

3) Intute

Intute 是英国 7 所大学合作构建的网络资源发现门户,它整合了 Altis、Artifact、BIOME、

EEVL、GSource、Humbul、PSigate、SOSIG 8 个非常有名的学科信息资源门户, 提供由学科馆员选择和评价的高质量教育和研究方面的英文网络资源。Intute 的前身是英国资源发现网络 (RDN, United Kingdom Resource Discovery Network), 是由英国联合信息系统委员会 (JISC, Joint Information Systems Committee) 资助、艺术与人文研究委员会及经济与社会研究委员会协助建设的大型项目。RDN 于 1999 年 10 月份正式启动, 2006 年 7 月更名为 Intute, 由曼彻斯特大学的 MIMAS 牵头, 众多合伙人和提供方共同协作, 整个组织的核心是一个包括 7 所大学在内的协会。目前已经发展成为由英国 70 多个教育和研究机构共同参与建设, 覆盖多类学科/专题领域的联合项目。

Intute 为英国的教育和研究团体免费提供发现和获取分布式、跨学科领域的高质量网络资源服务, 通过主题分类链接了 80 000 多个专业网站资源, 提供 10 万多个高质量教育和研究方面的网络资源链接服务。Intute 以提供高效的互联网资源来支持教育与科研活动, 目前共建立了社会科学类, 工程、数学与计算科学类, 健康与生命科学类, 物理科学类, 人文科学类, 工艺美术类, 休闲娱乐体育旅游类和地理环境类 8 个著名的学科信息资源门户, 分为科学技术、人文艺术、社会科学、健康与生命科学 4 个学科服务模块。

4) Infomine

Infomine 是为大学教师、学生和研究人员建立的网络学术资源虚拟图书馆, 其目的是为大学层次的教师、学生和研究人员提供相关的网络学术信息资源的导航服务。最初由加州大学图书馆于 1994 年 1 月创建, 目前的合作伙伴已包括加州大学所有院校及斯坦福大学等著名学府。Infomine 收藏有极为丰富的互联网资源, 它们包括各种重要的数据库、电子期刊、电子图书、公告板、讨论组、图书馆联机目录、教科书、会议论文集、研究人员的论文和名录及其他许多类型的信息。这些资源涉及几乎所有的学科领域, 仅生物科学领域 Infomine 就提供了可交互式检索的近 300 个数据库。

Infomine 对所有用户免费开放, 但是它提供的资源站点并不都是免费的, 能否免费使用取决于用户所在的图书馆是否拥有该资源的使用权。

7.5 搜索引擎

Internet 是一个广阔的信息海洋, 遨游其间而不迷失方向显得有些困难, 如何快速、准确地在网上找到所需信息已变得越来越重要。搜索引擎 (Search Engine) 作为人们访问 Internet、获取信息最主要的一种工具, 它能帮助网络用户在海量信息中迅速找到其所需要的信息资源。

7.5.1 搜索引擎概述

1. 搜索引擎的发展

1990 年以前, 没有人能够检索互联网上的信息。所有的网络信息检索工具都是从 1990 年蒙特利尔大学学生 Alan Emtage 等人发明 Archie 开始的, 虽然当时只实现了简单的 FTP 文件检索, 但它的工作原理与现在的搜索引擎很接近。随着 WWW 的出现和发展, 基于网页的搜索引擎于 1994 年 7 月出现, Michael Mauldin 将抓取程序 (Spider) 接入到索引程序中, 创建了大家熟知的 Lycos。同年 4 月, 斯坦福大学的两名博士生——David Filo 和美籍华人杨致远 (Gerry Yang) 共同创办了超级目录索引 Yahoo!, 并成功地使搜索引擎的概念深入人心, 从此搜索引擎进入高速发展时期。目前, 互联网上的搜索引擎已达数百家, 检索的信息量也与从前不可同日而语。

搜索引擎是一种能够通过 Internet 接收用户的查询指令,并向用户提供符合其查询要求的信息资源地址的系统。它在 Web 中主动搜索信息并将其索引的 Web 网站、其索引内容存储在可供检索的大型数据库中,提供索引和目录服务。

2. 搜索引擎的类型

网络搜索引擎是根据用户需求和资源特点建立的,网络用户需求的多样性和网络资源特点的多样性造成了搜索引擎类型的多样性。搜索引擎按照检索机制、检索范围、包含检索工具的数量、工作语种等特征可划分为多种类型。

1) 按检索机制划分

(1) 目录式搜索引擎。

目录式搜索引擎又称分类目录、主题指南,它按照某种分类体系组织网络资源,尤其是网站资源,提供一份按类别编排的 Internet 网站目录,各类目下排列着属于这一类别的网站站名和网址链接,有些还提供网站内容提要。

目录式搜索引擎虽有搜索功能,但在严格意义上说算不上是真正的搜索引擎,仅仅是按目录分类的网站链接列表而已。这类“搜索引擎”以人工方式或半自动方式搜集信息,由编辑或志愿者查看信息之后,人工形成信息摘要,并将信息置于事先确定的分类框架中。信息大多面向网站,提供目录浏览服务和直接检索服务。该类搜索引擎因加入了人的智能,所以信息准确、导航质量高;缺点是需要人工介入、维护量大、信息量少、信息更新不及时。Yahoo! 是目录式搜索引擎的典型代表,其他著名的目录式搜索引擎还有 Open Directory (DMOZ)、LookSmart 等,国内的搜狐、新浪、网易等门户网站也是从分类目录发展起来的。

(2) 全文搜索引擎。

全文搜索引擎是名副其实的搜索引擎,它通过信息抓取软件(如 Spider 或 Robot)从互联网上自动搜索网页,提取各网站的信息,建立索引数据库,提供关键词检索。用户通过检索界面输入检索词,计算机自动从索引数据库中匹配出相关记录,然后按一定的排列顺序将结果返回给用户。这种全文搜索引擎是真正意义上的搜索引擎,检索对象具体到网页。它的优点是资源覆盖率较高、数据库更新及时;缺点是基于全文索引和关键词简单匹配,检索准确性不高,噪声大、人工筛选负担重。代表性的全文搜索引擎主要有 Google、Bing、百度等。

目前,互联网信息检索工具,特别是一些综合性的检索工具多为混合型,既提供分类目录的浏览功能,又提供全文的关键词检索。如创立于 1998 年的搜索引擎 Google 以其简捷的关键词检索界面赢得了用户的青睐,短短几年发展成为全球最大、用户最多的综合性搜索引擎,但后来又将其 Open Directory 分类目录引入进来,提供网站分类目录浏览功能,成为混合型检索工具。

2) 按检索范围划分

搜索引擎按检索范围可划分为三种类型。

(1) 综合型搜索引擎。

综合型搜索引擎也称通用型搜索引擎,这类搜索引擎广泛搜集各种内容和媒体类型的资源,收录资源以广而全著称,主要面向互联网普通用户,满足一般信息需求。Google、Yahoo!、Bing、百度都属于这种类型。

(2) 专业型搜索引擎。

专业型搜索引擎主要面向专业用户,只收录一定专业领域的网络资源。这类搜索引擎对专业范围内资源的收录比较充分,信息加工质量也较高,往往具有一定的学术性,在资源加工和检索技术上较多地考虑专业使用的特点,检索效果较好。例如,一些计算机、医学、生物、法律领域的专业搜索引擎,数字图书馆或专业学科门户网站也属于这一范畴。

(3) 专门型搜索引擎。

专门型搜索引擎是查找特定类型信息的检索工具,如查找电话、电子邮件、地图、旅游信息、商业机构、图像、音乐等。这类搜索引擎一般要根据资源类型的特点,采用合适的信息描述和组织方式,提供特定的检索途径。目前,这类搜索引擎有的是一个独立的系统,如一些专业的地图搜索引擎、图像搜索引擎等;有的会集成到综合性搜索引擎中成为它的一个单元,如 Google 的地图搜索、图像搜索、视频搜索、博客搜索等。

3) 按包含检索工具的数量划分

按包含检索工具的数量不同,搜索引擎可分为独立搜索引擎和元搜索引擎。

(1) 独立搜索引擎。

这类搜索引擎建有自己的网络资源数据库,检索时只在自己的数据库内进行,由其反馈出相应的查询信息或链接的站点指向。每个独立的搜索引擎都会有自己的资源收录范围和查询特色,并据此提供相应的检索方式,如目录查询、全文查询、简单查询、高级查询等。独立搜索引擎是最常见的搜索引擎类型,各种常见的目录式搜索引擎和大多数全文搜索引擎都属于这一类型。

(2) 元搜索引擎。

元搜索引擎又称集成搜索引擎,它没有自己的网络资源数据库,而是将多个独立搜索引擎(源搜索引擎)集合在一起,提供一个统一的检索界面。当用户提出检索提问后,它会将其发送给多个搜索引擎,同时检索多个索引库,并根据一定规则进行相关度排序,将结果显示给用户。利用这类搜索引擎能够获得更大范围的信息源,检索的综合性、全面性也有所提高。著名的元搜索引擎有 MetaCrawler、Dogpile、Vivisimo、Mamma 等。从用户的角度来看,利用元搜索引擎的优点在于可以同时获得多个源搜索引擎的结果,但由于元搜索引擎在信息来源和技术方面都存在一定的限制,尤其是在高级检索和结果排序方面,因此搜索结果实际上并不理想。

4) 按工作语种划分

搜索引擎按工作语种可以区分为以下两种类型。

(1) 单语种搜索引擎。

单语种搜索引擎是指搜索时只能用一种语言查询的搜索引擎。

(2) 多语种搜索引擎。

多语种搜索引擎是指那些可以用多种语言或跨语言查询的搜索引擎,如 Google,支持 40 余种语言的查询。

5) 其他类型搜索引擎

(1) 桌面搜索引擎。

桌面搜索是 2004 年搜索引擎领域的热门词汇之一,已成为主要搜索引擎新的竞争领域,Google、Yahoo!、百度等纷纷推出各自的桌面搜索工具。与一般基于 Web 的搜索方式不同,桌面搜索的特点在于将搜索范围从互联网延伸到用户计算机硬盘中所存储的各种文档,如 E-mail、Word、Excel、PowerPoint、PDF 等多种格式的文本和音乐、图片和网页。换言之,先前的网络搜索引擎满足了用户搜索外界信息资源的需求,而桌面搜索则进一步帮助用户搜索定位自身已有的信息资源。面对庞大硬盘中日益膨胀、四处散落而难以统一管理利用的信息,桌面搜索引擎无疑很好地满足了用户快速便捷地管理、利用已有信息资源的需求。

(2) 垂直搜索引擎。

垂直搜索引擎是针对某一个行业的专业搜索引擎,是搜索引擎的细分和延伸。它通过对网页库中的某类专门信息的整合,定向分字段抽取出需要的数据,进行处理后再以某种形式返回给用户。垂直搜索引擎是针对通用搜索引擎信息量大、查询不准确、深度不够等弊端提

出来的一种新型搜索服务模式，它针对某一特定领域、某一特定人群或某一特定需求提供的有一定价值的信息及相关服务。其特点就是“专、精、深”，且具有行业色彩，是专门型搜索引擎的具体和深化，如国内提供旅游服务垂直搜索的酷讯网、汽车行业的新浪汽车搜索等。

（3）智能搜索引擎。

智能搜索引擎是结合了人工智能技术的新一代搜索引擎，是搜索引擎的主流发展方向。它除了能提供传统的快速检索、相关度排序等功能，还提供自然语言检索、用户兴趣自动识别、内容的语义理解、智能化信息过滤和信息推送等功能，能够根据用户的请求，从可以获得的网络资源中检索出对用户最有价值的信息。目前，真正实用的智能搜索引擎尚未出现，WolframAlpha (<http://www.wolframalpha.com/>)、微软的 Bing 搜索等已经具备了一些智能搜索的特征，是智能搜索引擎的初期产品。

7.5.2 搜索引擎工作机制

不管是目录式搜索引擎还是全文搜索引擎，它的工作机制一般由信息采集→信息组织→用户检索三大机制组成。

1. 目录式搜索引擎的工作机制

目录式搜索引擎将信息资源按照某种事先确定好的分类体系分门别类地逐层加以组织，用户通过浏览的方式层层遍历，逐层查找，直到找到所需要的信息线索，再通过信息线索链接到相应资源上。

目录式搜索引擎主要采用人工处理方式，以精选的网站或少量网页资源为对象，在人工描述、标引的基础上建立检索工具，以供用户浏览查找。这种网络信息检索工具比较适合那些“希望了解某一类信息，但又不严格限于查询关键词”的用户群。

其资源收集方式主要是人工采集和接受网站提交，同时也使用采集软件作为辅助手段，有针对性地采集资源，供人工处理。

网站分类专家将网络信息按照主题分成若干个大类，每个大类再分为若干个小类，依次细分，形成了一个可浏览的等级主题结构，即分类体系。一般的搜索引擎分类体系有五六层，有的甚至达十几层。编辑人员将收集来的网站或站点管理者提交来的网站进行分析审核，以决定是否收录；如果该站点审核通过，编辑人员还需要分析该站点的内容，并将该站点放在相应的类别和目录中。所有这些收录的站点被存放在一个“索引数据库”中。

用户在查询信息时，一般按照分类目录逐层地浏览查找，也可以选择按照关键词搜索，返回结果一般为网站，包括网站的名称、URL 地址、网站简介。如果以关键词搜索，则返回的结果跟全文搜索引擎一样，也是根据信息关联程度排列网站，但它的查询范围一般是网站的名称、网址、简介等，查询结果也只是网站首页 URL 地址，而非具体网页。

2. 全文搜索引擎的工作机制

全文搜索引擎是真正意义上的搜索引擎，它采用自动处理方式对网络信息资源进行采集、索引，并根据用户检索要求查找资源，将相关结果反馈给用户。它的基本结构仍由信息采集、信息组织和用户检索三部分组成，详细地说，由采集器、分析器、索引器、检索器和用户接口等五个部分组成，其工作机制图如图 7.20 所示。

1) 信息采集

信息采集工作由采集器和分析器共同完成，搜索引擎利用网络爬虫（Crawler）、网络蜘蛛（Spider）或网络机器人（Robot）等自动采集程序来获取网页上的超链接，通过请求 Web 站点上的 HTML 网页来采集该网页，遍历指定范围内的整个 Web 空间，不断从一个网页转

到另一个网页，从一个站点移动到另一个站点，将采集到的网页添加到网页数据库中。为保证采集的资料最新，还会定期回访已抓取过的网页，及时更新。

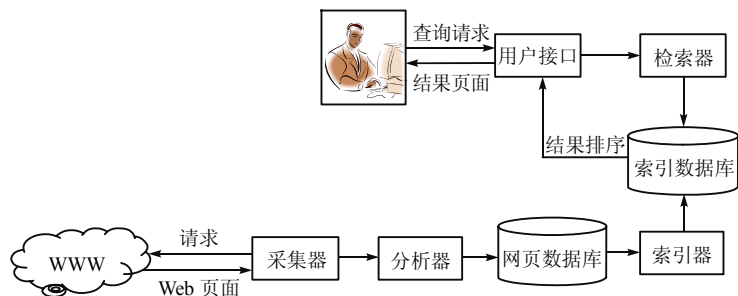


图 7.20 全文搜索引擎工作机制图

2) 信息组织

信息组织是搜索引擎为采集来的网页建立索引的过程，搜索引擎对采集来的网页进行组织，形成索引数据库，涉及网页结构分析、分词、排序等技术，好的索引能极大地提高检索速度。

3) 用户检索

用户向搜索引擎发出查询请求，搜索引擎根据用户需求到索引数据库中进行查找、匹配，并将检索到的结果以 Web 页面方式返回给用户。有的系统在返回结果之前对网页的相关度进行计算和网页等级评估，并根据相关度或网页等级进行排序，按逆序输出，因此，在不同的搜索引擎中搜索相同关键词，结果排序是不同的。

信息采集、信息组织与用户检索是搜索引擎工作最基本的三个组成部分，随着用户对搜索引擎性能要求的提高，不断有新的技术，如智能代理、用户相关反馈、搜索日志挖掘等技术应用到搜索引擎系统中，来完善搜索引擎的性能、增加搜索引擎的功能。

7.5.3 搜索引擎举要

1. Yahoo! (<http://www.yahoo.com>)

1994 年美国斯坦福大学的博士生 David Filo 和杨致远共同创立了被称为“网络指南信息库”的 Yahoo!，Yahoo! 是较早开发、以分类浏览而著称的搜索引擎。Yahoo! 目录有近 100 万个分类页面，14 个国家和地区当地语言的专门目录，包括英语、汉语、丹麦语、法语、德语、日语、韩语、西班牙语等。自问世以来，Yahoo! 分类目录已成为最常用的在线检索工具之一。

Yahoo! 是目录式搜索引擎的典范，它将网络资源划分为艺术人文、商业经济、计算机与互联网、教育、娱乐、政府、健康、新闻与媒体、科学、社会科学、社会与文化、休闲与运动、参考资料、国家地区等 14 个大类，提供分类式浏览服务。它的网站分类体系直到今日仍是其他搜索引擎纷纷效仿的对象。

Yahoo! 也提供关键词检索服务，分为基本检索和高级检索两种方式。基本检索是它的默认方式，用户可以在 Yahoo! 任意一个页面上找到其关键词检索输入框，在其中输入关键词，会返回相应的结果，并可以在 Web、Image、Video、Local、Shopping 等结果类型中自由切换，同时，Yahoo! 会推荐与检索词相关的词语供用户选择。Yahoo! 高级检索提供检索式的精化，允许用户使用与、或、非等逻辑操作，短语精确匹配，文件格式、语种、来源区域、

网站或网域等限定功能。

随着 Web 2.0 的发展, Yahoo! 还逐步确立了社区化搜索策略, 积极发动全球庞大的注册用户群来积累大批高质量网页内容和元数据, 从而改善用户的搜索体验。在这一策略下, Yahoo! 推出新的社区化搜索服务, 如知识堂等, 并收购了著名的照片共享网站 Flickr 和社会书签网站 Del.icio.us, 进行产品上的优势互补。

2. Google

Google (谷歌) (<http://www.google.com>) 由两个斯坦福大学博士生 Larry Page 与 Sergey Brin 于 1998 年 9 月创立。目前, Google 是全球用户最多、效果最佳的综合性搜索引擎之一。据 Google 官方博客宣称, Google 已经索引了一万亿网页, 每天提供亿次查询服务, 占全球搜索请求量的 1/3。

Google 基于引文分析原理的网页重要性评级技术 PageRank, 使得 Google 能自动将检索结果中大量低质量的网页过滤, 将高质量的权威网页排在检索结果的前面, 从而大大改进检索效果。

Google 最早以简捷界面的形式提供关键词检索服务, 也包括基本检索和高级检索, 其检索方式与其他搜索引擎相似。近年, Google 还引入了网景公司主持的大型公共网页目录 DMOZ, 推出了有史以来最翔实的网页目录。

专注检索是 Google 对搜索引擎网站最好的诠释。紧随网络发展和用户需求的变化, Google 陆续推出了学术搜索、地图搜索、图片搜索、图书搜索、生活搜索、桌面搜索、财经搜索、专利搜索、博客搜索等服务; 同时, Google 集成了网页快照、翻译、RSS 阅读等功能。

3. 百度

百度 (<http://www.baidu.com>) 于 1999 年在美国硅谷成立, 2000 年落户北京中关村, 是全球最大的中文搜索引擎, 目前中国提供搜索引擎服务的门户网站中 80% 以上由百度提供搜索引擎技术支持。百度专注中文搜索, 用户通过百度搜索引擎可以搜索到世界上最新、最全的中文信息。

百度采用超链分析技术来评价网站的质量, 这保证了用户在用百度搜索时, 越受用户欢迎的内容排名越靠前。

百度为中文用户度身定做, 根据中文用户搜索习惯开发出关键词自动提示功能、拼音检索功能、中文自动纠错功能等。

百度除了支持通用的简单检索和高级检索功能外, 还提供了多种特色搜索服务, 如文档搜索、MP3 搜索、视频搜索、图片搜索、新闻搜索、信息快递搜索、百度百科、百度知道、博客搜索、硬盘搜索等 50 余种特定类型搜索服务。

7.5.4 搜索引擎的发展方向

搜索引擎已成为人们访问互联网必不可少的一种工具, 随着“以人为本”的互联网服务理念深入, 搜索引擎将向着以下方向发展。

1. 智能化

搜索引擎的智能化方向发展是毫无疑问的。智能化检索是一种有效利用互联网信息的机制, 它使用自然语言处理、数据挖掘、机器翻译等人工智能技术实现信息采集智能化、信息组织智能化和用户检索智能化, 实现“口语化的提问, 智能化的结果”, 自动地将用户感兴趣的、对用户有用的信息以用户喜欢的方式提交给用户。

2. 桌面化

将搜索延伸到桌面是搜索引擎正在实现的一个方向,不仅是搜索软件桌面化,将搜索引擎作为工具栏嵌入到浏览器中;而且还是搜索资源的桌面化,将网络资源与本地资源整合,允许用搜索网络的方式来搜索本地硬盘中的各种资源,这实际上也是一种个性化,是搜索对象的个性化。

3. 细分化

搜索引擎细分化也称为垂直搜索,是针对具体行业、满足用户特定需求的专业搜索,是专业搜索引擎的细分和延伸。垂直搜索对某类专门信息进行结构化抽取和整合,以结构化检索方式提供精准的专业化检索结果。目前,如酷讯、房老大等生活服务类网站已提供了这种细分化的搜索服务,取得了良好的用户体验,细分化将是专业搜索引擎的主流发展方向。

4. 区域化

搜索区域化也称本地化,根据用户所在地区来限定搜索信息的范围,向用户优先推荐其所在地区的相关信息。据统计,向互联网搜索引擎提交的每4条搜索请求中,就有一条包含对本地信息的搜寻,预计这一比例还将上升。因此,Google、Yahoo!等搜索引擎公司不断推出各国、各地区的本地搜索网站或本地信息推送服务,搜索本地化已是必然趋势。

5. 个性化

个性化是“以用户为中心”的第二代互联网发展对搜索服务提出的要求。由于用户在知识背景和实际需求等方面的差异性,对于同一个检索需求,不同的用户会使用不同的描述方法,输入相同查询词的用户的信息需求也会大不相同。因此,搜索引擎还应该能根据不同用户的兴趣和具体情况,对用户查询做出不同的理解和语义的扩展,返回有针对性的检索结果以满足用户的不同需求。

6. 社交化

社会化网络的发展极大地影响了网络用户的信息行为,越来越多的用户愿意将自己对信息的感知通过转载、收藏、评论等方式分享给其他用户。搜索引擎社交化,能使用户获得来自好友、最信任信息源的信息推荐。Google、Bing等搜索引擎开始尝试与一些SNS网站合作,在搜索结果中添加社交和实时元素,使搜索结果能够更符合每个用户的搜索意图。

智能化是搜索引擎技术的发展方向,而桌面化、细分化、区域化都是个性化服务的内容,个性化、社交化是搜索引擎服务的发展趋向。



本章小结

目录、索引和文摘是应用最多的、使用范围最广的、综合性的信息组织工具。搜索引擎是网络信息资源组织的工具,也是网络资源组织的成果,它的出现和发展大大改变了人们的信息检索习惯,能够快捷、便利地满足大众用户的信息需求。全文数据库建设是对一次文献的数字化,既是信息资源的一种保存方式,也能实现资源的分布式分享和利用,对于资源保存、信息传播和学术研究具有重要意义。信息组织的目的是为了信息检索和利用,随着分布式资源的丰富,“以用户为中心”的网络信息服务理念要求信息提供商或服务商能够提供一站式整合信息平台,因此,异构数据库整合、学科信息门户被提出。由此可知,信息技术的变革和发展都是为了提供高质量信息服务、满足用户需求,信息组织与管理的目的也在于此。



问题讨论

1. 目录、索引和文摘各有什么特点？
2. 索引依据不同的划分标准可以划分为哪些类型？
3. 哪些信息技术的发展会影响搜索引擎的发展方向，未来的搜索引擎应该是什么样的？
4. 全文数据库的发展会不会取代传统的文摘数据库、题录数据库？在全文数据库日益强大的情况下，这样一些二次文献数据库该如何求生存？



第 8 章

语义网环境下的信息组织

内容提要

语义网的目标是通过给万维网上的文档添加能够被计算机所理解的语义 (Metadata), 让计算机能够“理解”分布在网上的信息和知识, 并在“理解”的前提下更好地处理、利用这些信息和知识。语义网技术可以引导人们进行语义层次上的信息分类、信息标引、信息整合等方式的信息组织, 实现一个有序的信息空间。资源描述框架 RDF (Resource Description Framework) 是语义网信息描述与表示的基础; 本体是语义网中信息组织的核心体系。

本章首先介绍了语义网中的信息描述与表示格式 RDF 数据模型、语义网中的信息建模方式本体及其构建方法以及语义网中的知识组织系统描述语言 SKOS, 然后举例说明在语义网环境下如何采用 OWL 本体对领域知识进行建模, 如何采用 RDF 语言基于本体对信息进行语义化的描述, 如何将描述好的信息在网络上发布为可访问的关联数据。

本章重点

- 语义网信息描述与表示;
 - 语义网信息组织模式;
 - SKOS (Simple Knowledge Organization System, 简单知识组织系统);
 - 语义网信息组织方法。
- 

8.1 语义网概述

自1991年诞生以来，Web（万维的简称）已经发展成为一个拥有亿级页面的巨大分布式信息空间，为用户提供海量的信息服务。但是当前Web还远远不能满足人们对信息共享和处理的需要，一方面人类对当前Web的利用无法得到软件工具的很好支持；另一方面Web上存在着大量信息孤岛，对Web上海量信息进行全局数据集成、信息共享、知识交换，并发现新知识，实现面向内容的信息管理，还存在许多困难，这主要归因于当前Web面向人类阅读和处理的特性。面对着当前Web在信息表达、组织、检索中存在的严重缺陷与不足，语义网应运而生。语义网的本质是采用一种适于机器理解和访问的新方式来表达Web上的内容，从而方便机器的阅读和处理，并在此基础上实现知识管理、概念检索、智能主体、普适计算等智能化的功能。

1998年9月，Web的缔造者Tim Berners-Lee在他的Web设计笔记里首次提出了语义网的设想^①。2000年12月，在XML2000会议上，Berners-Lee正式提出了下一代Web的概念——语义网（Semantic Web）及其体系结构，并于2001年5月在《科学美国人》杂志上发表了同名论文“The Semantic Web”，系统论述了他对下一代万维网架构语义网的蓝图，这篇论文同时也被认为是语义网诞生的标志。

语义网的目标是通过给Web上的文档添加能够被计算机所理解的语义（Meta data），让计算机能够“理解”分布在网上的信息和知识，并在“理解”的前提下更好地处理、利用这些信息和知识，从而使整个Web成为一个支持全球化知识共享的智能信息服务平台。由此看出，语义网相对于现有万维网的最大优势是“机器可理解”，它对Web的扩展可以使得Web具有知识理解及一定的推理和自动处理能力，它的出现给Web带来了革命性的变化，使人和机器协同工作理解并处理Web上的信息成为可能。

在XML2000大会上，Berners-Lee在综合了语义Web研究领域最新成果的基础上提出了语义网体系结构图（图8.1）。整个语义网的体系结构分为七层：XML（eXtended Markup Language）层为语法层（XML为标识语言）；RDF（Resource Description Framework）层为数据层（RDF为描述框架）；本体层（Ontology）为语义层（Semantic Layer，为词汇规范的语义网结构）；逻辑层（Logic）提供了智能推理的规则；证据层（Proof）支持代理间通讯的证据；确信层（Trust）和数字签名或加密技术则是为了保证信息交换的安全问题而设计的。在这个七层体系架构中，XML、RDF和本体是构建语义网的关键。

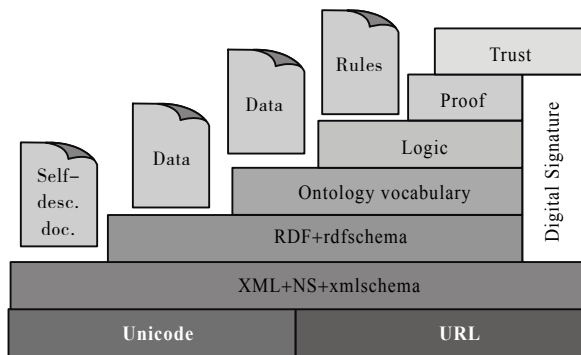


图 8.1 语义网标准体系架构

^① Berners-Lee, T. Design Issues: What the Semantic Web can represent [EB/OL]. [2014-03-03]. <http://www.w3.org/DesignIssues/RDFnot.html>.

8.2 语义网信息描述与表示

语义网信息描述与表示的基础是资源描述框架 RDF^① (Resource Description Framework)。RDF 是 W3C 于 1999 年提出的一种数据模型, 用于表示网络资源的元数据。

8.2.1 RDF 简介

RDF 定义了一个简单的数据模型, 通过主体 (Subject)、谓词 (Predicate)、客体 (Object) 这样的三元组来描述资源。RDF 是与语法无关的, 它可以建立在不同语法基础之上, 如可以通过图、三元组、自然语言、XML 等多种方法表示 RDF 模型。其中最重要的就是建立在 XML 语法之上的 RDF/XML 语法标准, 这是使用 RDF 表示 NKOS 的主要方式。图 8.2 就是用 RDF 描述资源的一个实例, 分别用图、三元组、XML、自然语言三种不同语法来描述同一个 RDF 资源模型, 这个模型中包含两个三元组:

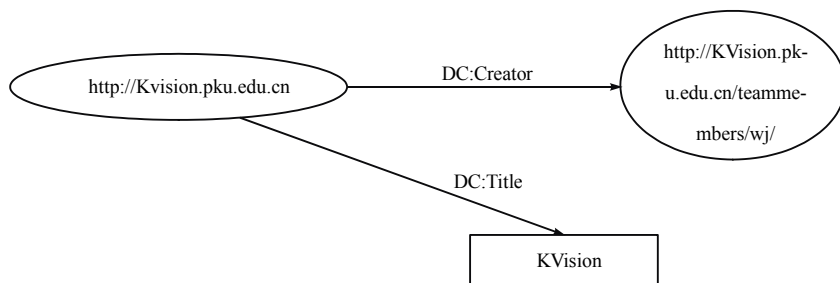


图 8.2 用 RDF 描述资源的一个实例

图 8.2 所示的 RDF 图使用三元组表示如下。

主体 (Subject)	谓词 (Predicate)	客体 (Object)
资源 (Resource) http://Kvision.pku.edu.cn	属性 (Property) DC:Creator	资源 (Resource) http://KVision.pku.edu.cn/teammembers/wj/
资源 (Resource) http://Kvision.pku.edu.cn	属性 (Property) DC: Title	文本 (Literal) “KVision”

使用 XML 表示为如下一段代码。

```

< ? xml version = "1.0"? >
< rdf:RDF
  xmlns:rdf = "http://www.w3.org/1999/02/22 - rdf - syntax - ns#"
  xmlns:DC = "http://purl.org/dc/1.1/">
  < rdf:Description about = "http://KVision.pku.edu.cn">
    < DC:Creator rdf:resource =
      "http://KVision.pku.edu.cn/teammembers/wj/" />
    < DC:Title > KVision < /DC:Title >
  < /rdf:Description >
< /rdf:RDF >
  
```

① RDF [EB/OL]. [2013-06-30]. <http://www.w3.org/RDF/>.

使用自然语言表达则是：资源“<http://KVision.pku.edu.cn>”的“DC:Creator”属性的值是“<http://KVision.pku.edu.cn/teammembers/wj/>”；资源“<http://KVision.pku.edu.cn>”的“DC:Title”属性的值是“KVision”。

RDF 的基本建模元语包括资源（Resource）、属性（Property）和语句（Statement）。其中资源可以充当 RDF 三元组的主体和客体，属性可以充当 RDF 三元组的谓词，而语句就是由 RDF 三元组本身组成的。

资源泛指所有能够用 RDF 表达式来表述的事物，如网页 <http://KVision.pku.edu.cn>；资源也可以是无法通过网络和 URI 来定位的对象，如现实中的一本书等；资源甚至可以是非物理存在的抽象概念，如“作者”这个概念。资源可以被 URI 引用（URIrefs，是一个在尾部附加了可选的“片断识别符”的 URI）标识。另外，RDF 中资源的集合可以形成容器（Container），容器包括三种：包（Bag）、序（Sequence）、替换（Alternative）；RDF 中还可以通过“集合”（Collection）来用列表结构表示一组资源，这个列表结构是用一些预定义的集合词汇表示的。关于容器和集合这里就不再详细介绍了。

属性是一种特殊的资源，用来描述资源的某个特定方面——通常是资源的元数据，如作者、标题等。属性可以是自定义的，也可以通过类似 XML 中 NS 机制来引用已定义的，如上例中 DC:Creator 就是引用已定义的都柏林核心集中的元素。

语句是由一个特定的资源、一个指定的属性及资源的这个属性的取值组成的，即 RDF 三元组的主体、谓词、客体共同组成一个语句。其中主体是资源，谓词是属性，客体可以是资源也可以是常量，如文本字符等。另外，语句可以通过复合形成高阶语句，如“资源‘<http://KVision.pku.edu.cn>’的‘DC:Title’属性的值是‘KVision’。”把上述句子看做一个资源，就会有下面的复合语句：“王老师说资源‘<http://KVision.pku.edu.cn>’的‘DC:Title’属性的值是‘KVision’。”为了能够表示这种高阶语句，RDF 使用了具体化（Reification）机制。

8.2.2 RDF 序列化表示格式

目前 RDF 有多种序列化格式，大致可分成四种类型：

（1）XML 类型：包括 RDF/XML、RDF/XML-ABBREV 和 Trix 三种格式，均采用 XML 编码，是面向机器阅读和处理的格式。RDF/XML 是 W3C 推荐的标准 RDF 序列化格式，RDF/XML-ABBREV 则是 RDF/XML 的压缩替代版，两者都遵从 RDF/XML 语法，在形式上没有太大区别。Trix，全称 Triples in XML，是一种用于命名图（Named Graphs）和 RDF 数据集（RDF Datasets）的序列化格式，为 RDF 图提供了一种规范、一致的 XML 表示。

（2）N3 类型：包括 N3、Turtle、N-Triples、N-Quads 和 TriG 格式，均采用纯文本表示，具有良好的可读性，是面向人类用户的格式。N3（Notation3）格式比 RDF/XML 格式更加紧凑、易读，但含有序列化 RDF 模型之外的一些特征（如支持基于 RDF 的规则）。Turtle 是 N3 的一个简化子集，删除了 N3 中与序列化 RDF 无关的规则，采用了省略与缩写，是一种适于手写的格式。N-Triples 又是 Turtle 的一个简化子集，抛弃了 Turtle 中的速记省略，以一行表示一个三元组，更加简洁易读。N-Quads 扩展了 N-Triples 格式，增加了一个可选的 context 值，即采用“<subject> <predicate> <object> <context>”四元组表示形式。TriG 则是基于 XML 的 Trix 格式的压缩替代版，同时也可看作是 Turtle 的一种轻度扩展。

（3）JSON 类型：包括 RDF/JSON 和 JSON-LD，均采用 JSON 兼容的格式表示。JSON（JavaScript Object Notation）是一种轻量级的数据交换语言，属于 Javascript 的一个子集，采用独立于语言的文本格式，既易于人的阅读，也易于计算机的解析。RDF/JSON 将一组 RDF 三元组表示为一系列嵌套的数据结构，JSON-LD 则是基于 JSON 的关联数据序列化格式。

(4) 嵌入式类型: 包括 Microformats、eRDF 和 RDFa, 均采用 XML 标签的形式将结构化的 RDF 三元组数据嵌入到 XHTML 网页中, 其目的是为了增强当前 Web 网页对 RDF 数据的支持。

以自然语言陈述“There is a Person identified by <http://www.w3.org/People/EM/contact#em>, whose full name is Eric Miller, whose email address is em@w3.org, and whose title is Dr.”为例, 不同的 RDF 序列化格式表示如下:

基于 XML 语法的 (RDF/XML) 表示:

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
  xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#em">
    <rdf:type rdf:resource="http://www.w3.org/2000/10/swap/pim/contact#Person"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#em">
    <contact:fullName>Eric Miller</contact:fullName>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#em">
    <contact:mailbox rdf:resource="mailto:em@w3.org"/>
  </rdf:Description>
  ....
</rdf:RDF>
```

Turtle 格式表示:

```
@prefix xsd:<http://www.w3.org/2001/XMLSchema#>.
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix contact:<http://www.w3.org/2000/10/swap/pim/contact#>.
<http://www.w3.org/People/EM/contact#em>
  rdf:type                contact:Person;
  contact:fullName        "Eric Miller"^^xsd:string;
  contact:mailbox         <mailto:fred@example.com>;
  contact:personalTitle   "Dr."^^xsd:string.
```

N-Triples 格式表示:

```
<http://www.w3.org/People/EM/contact#em>
  <http://www.w3.org/2000/10/swap/pim/contact#fullName>
    "Eric Miller"^^<http://www.w3.org/2001/XMLSchema#string>.

<http://www.w3.org/People/EM/contact#em> <http://www.w3.org/2000/10/swap/pim/contact#mailbox>
  <mailto:em@w3.org>

<http://www.w3.org/People/EM/contact#em>
  http://www.w3.org/2000/10/swap/pim/contact#personalTitle
  "Dr."^^<http://www.w3.org/2001/XMLSchema#string>.
```

8.2.3 RDF 评价

XML 只是一种语法规则，它本身无法表示机器可理解的语义，为此，W3C 推荐以 RDF 标准来解决 XML 的语义局限。

首先，RDF 希望以一种标准化，互操作的方式来规范 XML 的语义。XML 文档可以通过简单的方式实现对 RDF 的引用。通过在 XML 中引用 RDF，可以将 XML 的解析过程与解释过程相结合。也就是说，RDF 可以帮助解析器在阅读 XML 的同时，获得 XML 所要表达的主题和对象，并可以根据它们的关系进行推理，从而做出基于语义的判断。XML 的使用可以提高 Web 数据基于关键词检索的精度，而 RDF 与 XML 的结合则可以将 Web 数据基于关键词的检索更容易地推进到基于对象的检索。

其次，由于 RDF 是以一种建模的方式来描述数据语义的，这使得 RDF 可以不受具体语法表示的限制。但是 RDF 仍然需要一种合适的语法格式来实现 RDF 在 Web 上的应用。虽然 RDF 既可以用 N3 来表示，也可以用 XML 来表示。但是，由于 XML 已经成为被广泛支持的 Web 数据表示标准，便于应用的读取，因此将 RDF 序列化为 XML 表示可以使 RDF 获得更好的应用可处理特性，并使得 RDF 数据可以像 XML 数据一样的容易使用、传输和存储。

因此，RDF 和 XML 是互为补充的，而不只是对某个特定类型数据的规范表示，XML 和 RDF 的结合，不仅可以实现数据基于语义的描述，也充分发挥了 XML 与 RDF 的各自优点，便于 Web 数据的检索和相关知识的发现^①。当然，纵然 RDF 有如此优势，其语义表达能力依然非常有限：RDF 只提供了描述单个资源语义信息的能力，而没有提供描述特点领域的语义能力。因而 RDF 无法描述领域知识，无法抽象领域模型，所以，这还需要 RDF Schema 或者 OWL 等来进一步定义机器可理解的语义。

8.3 语义网信息组织模式

8.3.1 本体简介

本体是语义网中信息组织的核心体系，但它并不像哲学意义上那样抽象和理论化，是实实在在的信息描述的语言工具。

1993 年，Tom Gruber 将本体定义为“概念模型的明确的规范说明”，随后 Borst 和 Studer 等学者分别对此定义进行了进一步限定和阐述，将本体定义为“共享概念模型的明确的形式化规范说明”，这是当前较为普遍的关于本体概念的定义。

这个定义的具体含义由四个概念组成：

(1) “概念化”（Conceptualization），指将客观世界中的一些现象抽象出来得到的模型，它是客观世界的抽象和简化。

(2) “明确”（Explicit），即明确定义所使用的概念及概念的约束。

(3) “形式化”（Formal），即精确的逻辑表述，能够被计算机读取、理解和处理。

(4) “共享”（Shared），指本体描述的概念应该是某个领域公认的概念。

换言之，本体是对某一领域中的术语及术语间的关系做规范说明，提供对领域知识的共同理解和描述，用于共享、交流和复用，由经过精确定义的概念及概念间的关系组成，主要供计算机使用。

本体中的对象及它们之间的关系是通过知识表达语言的词汇来描述的。因此，可以通过

① RDF 和 XML 的区别[EB/OL].[2013-06-01].http://www.360doc.com/content/10/0426/10/865714_24925727.shtml

定义一套知识表达的专门术语来定义一个本体,来描述领域世界的实体、对象、关系及过程等,并通过形式化的公理来限制和规范这些术语的解释和使用。

根据 Perez 等人对本体建模的研究,本体可由以下五部分构成。

(1) 类 (Class): 类也称为概念,一般用于描述领域内具有相同属性或行为的一类对象的概念,如“人”是一个类,“教师”和“工程师”是“人”的子类,也是一个类。通过类的这种层级关系将本体中的概念组织成一个系统结构。

(2) 关系 (Relation): 关系是领域中类与类、实例与实例之间的联系,表示领域中概念或实例之间的交互作用,如 is-Child-of、a-Kind-of、IsA 等关系。

(3) 实例 (Instance): 实例是领域内某一特定的对象,根据本体的颗粒度的不同,对实例的界定也不同。“张三”是“教师”的实例,“李四”是“工程师”的实例,同时他们也都是“人”的实例,所以继承了“人”的各种属性。

(4) 函数 (Function): 一种特殊的关系,如 Mother-of 关系就是一个函数,其中, Mother-of (x, y) 表示 y 是 x 的母亲,显然 x 可以唯一确定他的母亲 y。

(5) 公理 (Axiom): 公理是领域内一些常识性知识的描述,是永真事实的描述。在应用本体的语义关系来进行逻辑推理的时候,这些规则能够发挥一定的作用。例如,“人是动物”就是公理。

此外,属性有时可以归于类,作为用来描述对象所具备性质的概念,有时可与关系划分到一起,是对概念内在特征和外在联系的揭示。函数和公理也可作为特殊的关系与关系归为一类。由上述描述可以看出,类、关系和实例是本体概念体系的基本构成,是领域对象概念化的主要元素。函数和公理的引入丰富了本体表达的内容,使得领域知识的表达变得更加严密和准确,并可通过推理发掘得出潜在的知识。

8.3.2 本体的类型

根据不同的分类标准,可以将本体分成多种类型。常用的本体主要有以下几种类型。

(1) 领域本体 (Domain Ontology): 领域本体是包含着特定领域概念、术语及关系的本体。该本体主要用于特定领域的应用。例如,经济类本体、建筑本体等。

(2) 通用本体 (Generic Ontology): 它覆盖了若干个领域,或者具有通用性,通常也被称为核心本体或顶级本体。它包括的是关于世界的一般性知识和概念,如时间、空间等。因此,通用本体可以跨学科领域使用,比较有代表性的通用本体当数 CYC。

(3) 应用本体 (Application Ontology): 应用本体是为某一特定的应用而建立的本体。例如,在某个数字图书馆的建设过程中,可以建立该图书馆的数字资源的本体,应用于该图书馆的信息表示与检索中。

(4) 任务本体 (Task Ontology): 任务本体描述的是特定任务或行为中的概念及概念之间的关系。任务本体与解决问题的方法相关,主要研究可共享的问题求解方法,这里的推理方法与领域无关,任务本体主要涉及动态知识,而不是静态知识,定义通用任务和推理活动,如诊断等。

另外,按照本体的形式化程度来说,本体可分为以下几种:① 完全非形式化本体;② 结构非形式化本体;③ 半形式化本体;④ 严格形式化本体。

8.3.3 本体的功能

本体对领域知识进行了一种表述,统一了领域内的术语和概念,便于人与机器、机器与机器之间的交流,从而增加知识共享、知识重用的程度。本体在信息组织方面的功能主要体

现在以下几个方面。

（1）信息描述方面，本体是关于领域知识的共同理解和描述，这使得基于本体的信息资源组织建立在语义层面而非语法层面，是以信息或知识的内容和本质特征为依据进行的组织。

（2）信息检索方面，本体具有良好的概念层次结构和对逻辑推理的支持，因而在信息检索，特别是在基于知识的语义检索中得到了广泛的应用。本体通过概念之间的关系来表达概念语义，因此，能实现基于本体的语义检索，避免当前信息检索因为字面检索而造成的低效率问题。

（3）语义网方面，本体面向计算机和网络的特点及形式化的描述使其能够更好地满足网络信息资源组织的需要，尤其是语义网信息组织的需要，所以它成为语义网体系框架中的一个主要层次。

本体能够准确地描述概念及概念之间的内在关联，并能通过逻辑推理获取概念之间蕴涵的关系，具有很强的表达概念语义和推理的能力，更适用于语义网环境中的信息组织和检索。

8.3.4 本体与传统情报检索语言的比较

本体与传统信息描述语言相比有许多相同之处，它继承了分类表的等级关系、叙词表的词汇控制等、本体语言与传统的信息描述语言相比，具有如下相似之处：

- （1）都是概念及概念关系的集合；
- （2）都是人们为便于实现人机或计算机之间的交流而制定的一致性标准，都能达到信息描述和提高信息检索效率的目的；
- （3）都可以看做是知识体系和结构的表现，都对词汇或概念实施了语义上的控制；
- （4）都适用于某一专业领域范围。

但本体也有许多不同于传统信息描述语言的地方，它具有面向计算机交流的特点，主要区别如表 8.1 所示。

表 8.1 本体语言与传统信息描述语言的比较

比 较 内 容	本 体	传统信息描述语言	
		主 题 语 言	分 类 语 言
概念模型	面向对象的认识世界的方法	面向概念的信息表示与检索方法	面向学科的信息表示与检索方法
组成元素	通常由类、属性、实例组成，有时包括函数和公理	语词及词间关系	类目及类目关系
标识	URI 唯一资源标识	语词	类号或类目
概念关系表达	几十种、上百种关系	等同、等级、相关三种关系	包含、并列、交替、相关等关系
形式化程度	较高	较低	较低
层级体系	存在，较为混乱，没有统一标准	有的存在，基本采用学科分类	存在，存在学科分类
适用对象	计算机	人为主，机器为辅	人为主，机器为辅
应用	提供语义检索和知识发现	信息内容的主题表示与检索	信息内容的分类表示与检索

8.3.5 本体的构建

1. 本体构建原则

本体构建的基本原则概括起来分为五大原则。

- （1）清晰（Clarity）：本体必须有效地说明所定义术语的意思。定义应该是客观的、与背

景独立的。当定义可以用逻辑公理表达时,它应该是形式化的。定义应该尽可能地完整。所有定义应该用自然语言加以说明。

(2) 一致 (Coherence): 本体应该是一致的,也就是说,它应该支持与其定义相一致的推理。它所定义的公理及用自然语言进行说明的文档都应该具有一致性。

(3) 可扩展性 (Extendibility): 本体应该为可预料到的任务提供概念基础。它应该可以支持在已有的概念基础上定义新的术语,以满足特殊的需求,而无须修改已有的概念定义。

(4) 编码偏好程度最小 (Minimal Encoding Bias): 概念的描述不应该依赖于某一种特殊的符号层的表示方法,因为实际的系统可能采用不同的知识表示方法。

(5) 本体约定最小 (Minimal Ontological Commitment): 本体约定应该最小,只要能够满足特定的知识共享需求即可。这可以通过定义约束最弱的公理及只定义通信所需的词汇来保证。

2. 本体开发工具

目前最流行的本体编辑工具是由美国斯坦福大学生物医学研究中心和医学院联合开发的开源软件 Protege 编辑器。Protege 有 3.X 和 4.X 两个版本。版本 3.X 已经基本稳定,最终版本是 Protege 3.4.5,版本 4.X 目前还在发展中,每隔一段时间就有新的版本推出,至 2011 年 4 月最新版本是 Protege 4.1。版本 3.X 支持 OWL 1.0 和 RDFS,版本 4.X 只支持 OWL 2.0。Protege 还有许多插件可以下载,对其功能进行了扩展,譬如图形显示、SKOS 编辑器、推理器等。

除 Protege 之外,还有其他的本体编辑工具,如斯坦福大学知识系统实验室的 Ontolingua 和 Ontoprise GmbH 公司的商业软件 OntoStudio 等。2010 年 2 月欧盟研究项目 NeOn 发布了免费的本体工具包 NeOn Toolkit 2.3,该工具包很大程度上是基于商业软件 OntoStudio,但进行了进一步的扩展。

3. 本体构建方法

国内外在本体构建方法上,研究主要是从知识工程的角度,探讨本体的构建方法,也称为本体工程。

本体工程的主要特点是强调构建本体时要按照一定的规范和标准。到目前为止,本体工程中比较有名的几种方法包括:TOVE 法、METHONTOLOGY 法、骨架法、KACTUS 法、SENSUS 法、DEF5 法和七步法等。其中以骨架法和七步法应用最为广泛。

(1) 骨架法。

骨架法 (Skeletal Methodology) 由 Mike Uschold 和 Micheal Gruninger 提出,又称 Enterprise 法,专门用来创建企业建模过程中的本体。骨架法流程如图 8.3 所示。

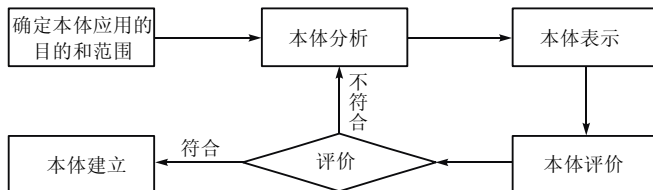


图 8.3 骨架法流程

① 确定本体应用的目的和范围: 根据所研究的领域或任务,建立相应的领域本体或任务本体,领域越大,所建本体越大,因此需限制研究的范围。

② 本体分析: 定义本体所有术语的意义及其之间的关系,该步骤需要领域专家的参与,

对该领域越了解，所建本体就越完善。

③ 本体表示：以本体表示语言对本体进行描述。

④ 本体评价：按照对本体表示的清晰性、一致性、完整性、可扩展性，对所建立的本体进行评价，如果符合要求则进入本体建立，不符合要求则返回第二步重新进行本体分析。

⑤ 本体的建立：对所有本体按以上标准进行检验，符合要求的以文件形式存放，生成 OWL、RDF 等格式的本体文件。

(2) 七步法。

七步法由斯坦福大学医学院开发，主要用于领域本体的构建。七个步骤分别是：

① 确定本体的专业领域和范畴。领域知识往往十分庞大，本体不可能包括所有的概念，因此，在建立本体之前必须先确定本体将覆盖的专业领域、范围和应用目标等。不同的应用领域，领域概念肯定不同；即使是同一个领域，由于应用目的不同，本体所包容的概念也会有所不同。因此，建立本体之前一定要明确本体建立的领域、范畴和应用目标。

② 考查复用现有本体的可能性。共享和复用是本体的特点，建立本体的目的也是为了解决知识的共享和复用问题，因此在设计和建立本体之前，应该考虑是否有已经建立好的本体供复用。

③ 列出本体中的重要术语。领域本体是描述概念及概念与概念之间的关系的，首先要列举出该领域中的所有概念及对该概念的详细解释。在特定领域，这些概念就是与领域相关的专业术语。把领域中一些重要术语列举出来，有利于本体工程师更好地理解本体建立的目标，明确方向。

④ 定义类和类的等级体系，通常采用自顶向下法 (Top-down)、自底向上法 (Bottom-up) 和综合法。通过等级体系将领域概念进行分类组织，用于描述领域概念间的类属关系，并将本体中的概念模块化。本体中的概念模块化应包括：概念名称、语义描述、概念可能的同义词、缩略语。定义分类概念，就是对这些信息进行描述；同时，对所建立的概念分类层次进行检验，保证没有重复的概念，防止冗余定义。

⑤ 定义类的属性。概念的分类层次结构体现了分类概念之间的一种继承关系 (Kind-of)，但是在领域本体中，概念和概念之间通过关系来交互，除了继承关系，在我们构建的领域本体中还可以根据需要定义其他的关系。针对每个概念，要列出它所有可能的属性，每个属性都有对应的属性值。

⑥ 定义属性的分面。属性的分面可用来描述属性值的类型、允许的取值、取值的个数 (基数)，以及属性其他的一些特征。

⑦ 创建实例。确定类的实例首先需要选取一个概念类，然后添加属于该类的具体实例，最后为实例添加具体属性值。

4. 本体构建实例

在本节中，参考斯坦福大学医学院七步法本体构建原则，基于 DC 元数据标准构建一个核心元数据本体（命名空间前缀为 co）。核心元数据本体是一个描述文献资源的基本模型，定义了文献资源的核心属性、文献资源之间以及文献资源与其他相关资源（如知识组织资源、个人、组织机构、地点）之间的基本关系。

DC 是一套描述网络资源的元素集合。基于 DC 元数据标准，该元数据本体完全复用 DC 元数据的 15 个元素，包括：Title、Creator、Date、Subject、Publisher、Type、Description、Contributor、Format、Source、Rights、Identifier、Language、Relation、Coverage。其构建具体步骤如下：

(1) 元数据元素分析。

元数据的描述对象是文献资源，因此在元数据本体中，最核心的类应是“文献资源”，

标识为“co:Document”，是图书馆中各种文献资源对象的集合。该类还可以包含多个子类，代表各种不同的文献资源类型。

15个核心元素（即DC元数据）的取值类型可以分为四类：自由文本、时间日期、实体对象、来自受控词表的规范术语。

（2）定义具名类（Named Classes）。

所谓类是一组具有共同属性的个体（或对象）的集合。类分为具名类（Named Classes）和匿名类（Anonymous Classes）。

在构建核心元数据本体时，首先需要根据元数据的描述对象和元数据元素的取值范围定义本体中的具名类。其中自由文本直接采用字符串数据类型表示，而其他三种取值范围则需要定义类来表示。因此定义一个表示时间日期集合的类 co:PeriodOfTime；定义一个表示个人和组织机构实体对象集合的类 co:Agent；另外直接采用 SKOS 模型中定义的 skos:Concept Scheme 类表示受控词表的集合，采用 skos:Concept 类表示受控词表中规范术语的集合。

下面是对核心元数据本体中定义的几个具名类的详细描述。

① 定义 co:Document（文献资源）类。

该类代表的是图书馆中各种文献资源的总集合，然后可以为该类构建各种子类，构成一个层次化的分类体系，代表不同类型的文献资源。关于“文献资源”的分类有多种途径和标准，建议参照《国家图书馆核心元数据标准著录规则》^①中所附的“信息资源名称规范列表”和 Bibliographic Ontology^②中定义的文献资源分类体系进行分类。表 8.2 给出了 co:Document 类的一个不完全分类层次作为示例和参考。

表 8.2 核心元数据本体中 co:Document 类的部分子类

第二层级类	第三层级类	第四层级类
co:Document (文档)	co:Ancient（古籍文献）*	co:RareBook（古籍图书）
		co:Chorography（地方志）
		co:Rubbing（拓片）
		co:Genealogy（家谱）
		co:Atlas（舆图）
	co:Book（图书）	co:Proceeding（会议录）
	co:Article（文章）	
	co:Image（图像）	co:Map（地图）
		co:Photo（照片）
		co:Picture（书画）
	co:Audio（音频文件）	
	co:Video（视频文件）	
	co:Manuscript（手稿）	
	co:Patent（专利文献）	
	co:Report（科技报告）	
	co:Standard（技术标准）	
	co:Thesis（学位论文）	

① 沈芸芸, 裴微微. 国家图书馆核心元数据标准著录规则（初稿）[R]. 国家图书馆, 2009, 9.

② D'Arcus, B. & Giasson, F. Bibliographic Ontology Specification [EB/OL]. [2010-04-28].
http://bibliontology.com/specification

续表

第二层级类	第三层级类	第四层级类
co:Document (文档)	co:Reference (参考资源)	
	co:Slideshow (幻灯片)	
	co:Webpage (网页)	
	co:PersonalCommunicationDocument (个人交流文档)	co:Email (电子邮件) co:Letter (书信)
	co:Website (网站)	
co:Collection (文档集合)	co:Series (丛书)	
	co:Periodical (定期出版物)	co:Newspaper (报纸)
		co:Magazine (杂志)
		co:Journal (期刊)
co:Component (文档部件)	co:BookSection (书的组成部分)	co:Chapter (章节)
	co:Excerpt (摘录)	co:Quote (引用)
	co:Slide (单张幻灯片)	

注:

(1) 英文名称是类的本地标识符, 扩号中是类的中文标签;

(2) co 代表核心元数据本体的命名空间 “http://hostname/onto/core”

② 定义 co:Agent (代理者) 类。

该类代表的是个人、团体、组织机构以及服务的集合, 与 FOAF 本体中的 foaf:Agent 类相等价, 也与 DCMI 元数据术语中的 dcterms:Agent 类相等价。因为在 FOAF 本体中已经对 foaf:Agent 类进行了详细的定义和描述, 完全可以借用这些定义和描述以用于 co:Agent 类。co:Agent 类含有三个不相交的子类, 与 foaf:Agent 类的三个相应子类完全等价:

- co:Agent owl:equivalentClass dcterms:Agent
- co:Agent owl:equivalentClass foaf:Agent
 - co:Person owl:equivalentClass foaf:Person (个人)
 - co:Organization owl:equivalentClass foaf:Organization (组织机构)
 - co:Group owl:equivalentClass foaf:Group (团体)

③ 定义 co:PeriodOfTime (时间) 类。

该类代表各种时间日期表示的集合。时间日期表示可以分为两大类: 一类是表示一个时间点, 譬如某年、某月、某日; 另一类是表示一个时间段, 由起始时间点和结束时间点来限定。因此 co:PeriodOfTime 类含有两个子类:

- co:PeriodOfTime
 - co:Instant (表示时间点)
 - co:Interval (表示时间段, 由起始时间点和结束时间点来限定)

④ 复用 skos:ConceptScheme (概念体系) 类。

该类是 SKOS 核心模型中的定义的一个类, 代表概念框架的集合。在 SKOS 模型中, 一个受控词表可以描述为一个 SKOS 概念框架, 即 skos:ConceptScheme 类中的一个实例。如果几个受控词表描述的概念相似或相近 (譬如均描述地名概念), 可以合并在一个 SKOS 概念框架中进行描述。表 8.3 列出了《国家图书馆核心元数据标准著录规则》中所用到的受控词表以及与之相应的 URI 标识符。

表 8.3 《国家图书馆核心元数据标准著录规则》中用到的受控词表及其 URI 标识符

元数据元素	取值的受控词表	SKOS 语义化受控词表的 URI 标识符
language	RFC 4646	http://www.nlc.gov/vocab/RFC4646
	ISO 639-2	http://www.nlc.gov/vocab/ISO639-2
	世界语种代码表（国家图书馆编）	http://www.nlc.gov/vocab/Lang
	中国少数民族文字表（北京大学图书馆编）	http://www.nlc.gov/vocab/EthnicLang
subject	中国图书馆分类法（CLC）	http://www.nlc.gov.cn/vocab/CCT_CLC_v4.0
	汉语主题词表（CT）	http://www.nlc.gov.cn/vocab/CCT_CT_v2.0
	古籍四部分类法（FDC）	http://www.nlc.gov.cn/vocab/FDC
	中科院图书馆图书分类法（LASC）	http://www.nlc.gov.cn/vocab/LASC
type	DCMType 类型词汇表	http://www.nlc.gov.cn/vocab/DCMType
	信息资源名称规范列表	http://www.nlc.gov.cn/vocab/IRType
format	Internet 媒体类型 IMT（MIME）	http://www.nlc.gov.cn/vocab/IMT
	CALIS 采用的规范表	— —
	科技部项目中的数字资源格式	— —
coverage. spatial	地理名称叙词表（TGN）	http://www.nlc.gov.cn/vocab/TGN
	地理科学叙词表（TGS）	http://www.nlc.gov.cn/vocab/TGS
	ISO 3166（国家名称代码）	http://www.nlc.gov.cn/vocab/ISO3166
	中国图书馆分类法地区复分表	http://www.nlc.gov.cn/vocab/CCT_CLC_Place_v4.0

⑤ 复用 skos:Concept（概念）类。

该类是 SKOS 核心模型中定义的一个类，代表概念的集合。在 SKOS 核心模型中，受控词表中的一个规范术语可以描述为一个 SKOS 概念，即 skos:Concept 类的一个个体。概念和概念框架之间的关系用 skos:hasTopConcept 和 skos:inScheme 属性来描述。在该语义化描述规范中对 SKOS 核心模型进行了扩展，扩展后的模型被称作 SKOSEX 模型，是针对中文网络知识组织系统进行语义化描述的一个模型。在 SKOSEX 模型中定义了 skos:Concept 类的一系列子类用以表示特定类型的 SKOS 概念，如下所示：

- skos:Concept
 - skosex:LanugageConcept （表示语种概念的集合）
 - skosex:LocationConcept （表示地名概念的集合）
 - skosex:ResourceTypeConcept （表示资源类型概念的集合）
 - skosex:MediaFormatConcept （表示媒体类型概念的集合）
 - skosex:PersonConcept （表示人名概念的集合）
 - skosex:OrganizationConcept （表示团体名和组织机构名概念的集合）

来源于某些受控词表的概念可能全部属于某个特定的子类，如来源于语种列表的概念全部属于 skosex:LanugageConcept 类，但是也有些受控词表中的概念非常多样，属于整个 skos:Concept 类而不是某个特定的子类，如“汉语主题词表”中的概念。表 8.4 列举了《国家图书馆核心元数据标准著录规则》中所用到的受控词表中以及这些受控词表中的概念所属的 skos:Concept 类的子类。

表 8.4 《国家图书馆核心元数据标准著录规则》中用到的受控词表及其概念所属的 SKOS 概念类型

受 控 词 表	skos:Concept 类或其子类	注 释
<ul style="list-style-type: none"> ➤ RFC 4646 ➤ ISO 639-2 ➤ 世界语种代码表 ➤ 中国少数民族文字表 ➤ 其他语种词表 	skosex:LanguageConcept	语种概念
<ul style="list-style-type: none"> ➤ 汉语主题词表（CT） ➤ 其他叙词表 ➤ 中国图书馆分类法（CLC） ➤ 中国科学院图书分类法（LASC） ➤ 古籍四部分类法（FDC） ➤ 其他分类法 	skosex:Concept	叙词表和分类法中的概念非常多样，因此属于整个 skos:Concept 类，而非特定的子类
<ul style="list-style-type: none"> ➤ DCMI 类型词汇表 ➤ 信息资源名称规范列表 ➤ 其他资源类型表 	skosex:ResourceTypeConcept	资源类型概念
<ul style="list-style-type: none"> ➤ IMT（互联网媒体类型）词表 ➤ CALIS 采用的规范表 ➤ 科技部项目中的数字资源格式 ➤ 其他媒体类型表 	skosex:MediaFormatConcept	媒体类型概念
<ul style="list-style-type: none"> ➤ 中国图书馆分类法地区复分表 ➤ 地理名称叙词表（TGN） ➤ 地理科学叙词表（TGS） ➤ ISO 3166（国家名称代码） ➤ 汉语主题词表中的地名表 ➤ 其他地名词表 	skosex:LocationConcept	地名概念
<ul style="list-style-type: none"> ➤ 人名/组织机构名规范档 	skosex:PersonConcept skosex:OrganizationConcept	个人、团体和组织机构名概念

注：

(1) skos 代表 SKOS 核心模型的命名空间 “<http://www.w3.org/2004/02/skos/core#>”；

(2) skosex 代表 SKOS 扩展模型的命名空间 “<http://www.w3.org/2004/02/skos/extension#>”

图 8.4 给出了一个采用 SKOS 语言描述受控词表中规范术语的示例。“ISO 639-2 语种列表”和“世界语种代码表”被分别描述为一个概念框架（skos:ConceptScheme）。“ISO 639-2 语种列表”中的规范术语“Chinese”和“世界语种代码表”中的规范术语“汉语”被分别描述为相应概念框架中的一个 SKOS 概念（skos:Concept）。位于不同概念框架中的这两个概念含义完全相同，因此是精确匹配（skos:exactMatch）的关系。图 8.4 所示的是上述描述的 Turtle 序列化表示。


```

@prefix iso: <http://www.nlc.gov.cn/vocab/ISO639-2#>.
@prefix lang: <http://www.nlc.gov.cn/vocab/Lang#>.
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix skosex: <http://www.w3.org/2004/02/skos/extension#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
<http://www.nlc.gov.cn/vocab/ISO639-2>    rdf:type    skos:ConceptShceme.
<http://www.nlc.gov.cn/vocab/Lang>        rdf:type    skos:ConceptShceme.
iso:chi    rdf:type          ckos:LanguageConcept;
            skos:inScheme     <http://www.nlc.gov.cn/vocab/ISO639-2>;
            skos:notation     “chi” ^^<Notation1>;
            skos:notation     “zh” ^^<Notation2>;
            skos:prefLabel    “Chinese” @en;
            skos:prefLabel    “Chinois” @fr;
            skos:exactMatch   lang:chi .
lang:chi    rdf:type          ckos:LanguageConcept;
            skos:inScheme     <http://www.nlc.gov.cn/vocab/Lang>;
            skos:notation     “chi” ^^<Notation1>;
            skos:prefLabel    “Chinese” @en;
            skos:prefLabel    “汉语” @zh;
            skos:exactMatch   iso:chi.

```

图 8.4 采用 SKOS 语言描述的语种词表中术语的语义化描述示例 (Turtle 格式)

(3) 定义属性 (Properties)。

在定义好本体的类之后, 需要定义属性来描述类的特征以及类与类之间的关系。本体的两种主要属性类型: 数据类型属性和对象属性, 都具有领域 (Domain) 和值域 (Range)。如果是数据类型属性, 则值域是字符串、数字、时间日期等文字值; 如果是对象属性, 则值域是某个 (些) 类。属性的实质就是描述领域中的个体与值域中的个体之间的相互关系。

① 领域为 co:Document 类的属性

《国家图书馆核心元数据标准》^①中定义的 15 个元数据元素就是为描述文献资源的属性而制定的, 因此制定 “co:Document” 类的属性的依据就是这 15 个元数据元素。其中大多数元素可以直接复用为本体的属性, 但是有些元素需进行分解, 转换为多个属性。领域为 co:Document 类的所有属性如表 8.5 所示, 具体说明如下。

表 8.5 核心元数据本体中 co:Document 类的属性

属性标识符	标 签	属 性 类 型	值 域	出 处
dc:title	名称	数据	xsd:string	DC
dc:description	描述	数据	xsd:string	DC
dc:identifier	标识符	数据	xsd:string	DC
dc:date	日期	对象	co:PeriodOfTime	DC
dc:creator	创建者	对象	co:Agent 或 foaf:Agent 或	DC
dc:contributor	其他责任者	对象	skosex:PersonConcept+ skosex:OrganizationConcept	DC
dc:publisher	出版者	对象		DC

① 沈芸芸, 裴微微. 国家图书馆核心元数据标准 (1.0 版) [R]. 国家图书馆, 2009, 9.

续表

属性标识符	标 签	属 性 类 型	值 域	出 处
dc:relation	关联	对象	co:DocumentResource	DC
dc:source	来源	对象	co:DocumentResource	DC
dc:language	语种	对象	skosex:LanguageConcept	DC
dc:subject	主题	对象	skos:Concept	DC
dc:type	类型	对象	skosex:ResourceTypeConcept	DC
dc:rights	权限	数据	xsd:string	DC
dcterms:temporal	时间范围	对象	co:PeriodOfTime	DC Terms
dcterms:spatial	空间范围	对象	skosex:LocationConcept 或 co:Location	DC Terms
dc:format	格式	数据	xsd:string	DC
co:media	媒体类型	对象	skosex:MediaFormatConcept	新定义

注：

- (1) dc 代表 DC Metadata Element Set 1.1 的命名空间 “http://purl.org/dc/elements/1.1/”；
- (2) dcterms 代表 DCMI Metadata Terms 的命名空间 “http://purl.org/dc/terms/”；
- (3) co 代表核心元数据本体的命名空间 “http://hostname/onto/core”；
- (4) dc:identifier 的属性值是除 URI 标识符以外的其他标识符；
- (5) skosex: 代表 SKOS 的扩展部分

② 领域为 co:PeriodOfTime 类的属性。

co:PeriodOfTime 类包含两个子类：co:Instant 子类和 co:Interval 子类。co:Instant 子类表示时间点，因此它的属性是时间点的日、月、年或者年号纪年表示值。co:Interval 表示时间段，因此它的属性是时间段的起始时间点和结束时间点的日、月、年或者年号纪年表示值。文献资源的时间日期表示都是以日期和年代表的粗粒度时间，因此 co:PeriodOfTime 类不含有表示时、分、秒等细粒度时间的属性。核心元数据本体中领域为 co:PeriodOfTime 类的所有属性如表 8.6 所示。

表 8.6 核心元数据本体中 co:PeriodOfTime 类的属性

属性标识符	属 性 类 型	标 签	领域 (Domain)	值域 (Range)	注 释
co:dayValue	数据	日	co:Instant	xsd:gDay	格式为---DD，如---14
co:monthValue	数据	月	co:Instant	xsd:gMonth	格式为，如—MM，如--04
co:yearValue	数据	年	co:Instant	xsd:gYear	格式为 CCYY，如 2010
co:eraValue	数据	纪年	co:Instant	xsd:string	年号纪年字符串，如清康熙 25 年
co:startDayValue	数据	起始日	co:Interval	xsd:gDay	如---01
co:endDayValue	数据	结束日	co:Interval	xsd:gDay	如---10
co:startMonthValue	数据	开始月	co:Interval	xsd:gMonth	如--01
co:endMonthValue	数据	结束月	co:Interval	xsd:gMonth	如--10
co:startYearValue	数据	起始年	co:Interval	xsd:gYear	如 2001
co:endYearValue	数据	结束年	co:Interval	xsd:gYear	如 2010
co:startEraValue	数据	起始纪年	co:Interval	xsd:string	以年号表示的起始年，如民国元年
co:endEraValue	数据	结束纪年	co:Interval	xsd:string	以年号表示的结束年，如民国 11 年

③ 领域为 `co:Agent` 类的属性。

CO 核心元数据本体中的 `co:Agent` 类与 FOAF 本体中的 `foaf:Agent` 类相等价, 因此 FOAF 本体^①中定义的 `foaf:Agent` 类的属性可以完全用于 `co:Agent` 类, 如表 8.7 所示。如果需要, 还可以对 FOAF 本体进行扩展, 定义新的属性作为补充。

表 8.7 FOAF 本体中 `foaf:Agent` 类的主要属性

属性标识符	属 性 类 型	领 域 (domain)	值 域 (range)	注 释
<code>foaf:name</code>	数据	<code>owl:Thing</code>	<code>xsd:string</code>	名称
<code>foaf:title</code>	数据	<code>foaf:Person</code>	<code>xsd:string</code>	个人的头衔
<code>foaf:mbox</code>	对象	<code>foaf:Agent</code>	<code>owl:Thing</code>	代理者的邮箱
<code>foaf:phone</code>	对象	暂无定义	暂无定义	电话信息
<code>foaf:birthday</code>	数据	<code>foaf:Agent</code>	<code>xsd:date</code>	代理者的出生/产生日期, 格式 MM-DD
<code>foaf:gender</code>	数据	<code>foaf:Agent</code>	<code>xsd:string</code>	代理者的性别
<code>foaf:holdsacount</code>	对象	<code>foaf:Agent</code>	<code>foaf:OnlineAccount</code>	代理者持有的网络帐号
<code>foaf:knows</code>	对象	<code>foaf:Person</code>	<code>foaf:Person</code>	一个人认识的其他人
<code>foaf:schoolHomepage</code>	对象	<code>foaf:Person</code>	<code>foaf:Document</code>	个人学习过的院校的主页
<code>foaf:workInfoHomepage</code>	对象	<code>foaf:Person</code>	<code>foaf:Document</code>	个人的工作信息主页
<code>foaf:workPlaceHomepage</code>	对象	<code>foaf:Person</code>	<code>foaf:Document</code>	个人所在机构的主页
<code>foaf:img</code>	对象	<code>foaf:Person</code>	<code>foaf:Image</code>	个人的图片
<code>foaf:topic_interest</code>	对象	<code>foaf:Person</code>	<code>owl:Thing</code>	个人感兴趣的主体
<code>foaf:interest</code>	对象	<code>foaf:Person</code>	<code>foaf:Document</code>	个人感兴趣的主体的文档
<code>foaf:publications</code>	对象	<code>foaf:Person</code>	<code>foaf:Document</code>	个人的出版物
<code>foaf:pastProject</code>	对象	<code>foaf:Person</code>	<code>owl:Thing</code>	个人从事过的项目
<code>foaf:currentProject</code>	对象	<code>foaf:Person</code>	<code>owl:Thing</code>	个人正从事的项目
<code>foaf:member</code>	对象	<code>foaf:Group</code>	<code>foaf:Agent</code>	团体的成员
<code>foaf:made</code>	对象	<code>foaf:Agent</code>	<code>owl:Thing</code>	代理者所做过的事情
<code>foaf:weblog</code>	对象	<code>foaf:Agent</code>	<code>foaf:Document</code>	代理者的博客

注:

- (1) 属性 `foaf:topic_interest` 的值域可缩小到 `skos:Concept` 类, 即采用 SKOS 概念描述个人感兴趣的主体。
- (2) FOAF 本体中的 `foaf:Document` 类与核心元数据本体中的 `co:Document` 类等价。

④ 领域为 `skos:ConceptScheme` 和 `skos:Concept` 类的属性。

`skos:ConceptScheme` 类和 `skos:Concept` 类都是 SKOS 核心模型中定义的类, 它们的属性也已经在 SKOS 核心模型中进行了详细定义, 详细内容请参见《SKOS Simple Knowledge Organization System Primer》^②。

(4) 设定逆属性 (Inverse Properties)

本体中每个对象属性都可能有一个相应的逆属性。所谓逆属性, 是指如果属性 `p` 连接两

① Brickley, D. & Miller, L. FOAF Vocabulary Specification 0.97 (Namespace Document 1 January 2010 - 3D Edition) [EB/OL]. [2010-04-28]. <http://xmlns.com/foaf/spec/>.

② Isaac, A., & Summers, E. SKOS Simple Knowledge Organization System Primer (W3C Working Group Draft 21 February 2008) [EB/OL]. [2011-04-10]. <http://www.w3.org/TR/skos-primer/>.

个个体 A 与 B ($p: A \rightarrow B$), 那么连接 B 与 A ($q: B \rightarrow A$) 的属性 q 就是属性 p 的逆属性。一对互逆属性的领域和值域正好相互交换。下列属性对互为逆属性:

- dcterms:hasPart owl:inverseOf dcterms:isPartOf
- dcterms:hasFormat owl:inverseOf dcterms:isFormatOf
- dcterms:hasVersion owl:inverseOf dcterms:isVersionOf
- dcterms:replaces owl:inverseOf dcterms:isReplacedOf
- dcterms:requires owl:inverseOf dcterms:isRequiredBy
- dcterms:references owl:inverseOf dcterms:isReferencedBy

(5) 设定属性特征 (Characteristics)。

属性特征在 CO 本体中的应用并不显著。但是在表示文献资源个体之间复杂的书目关系时比较有用, 这里进行简单的介绍。数据类型属性只有函数属性这一个特征, 对象属性则有四个特征: 函数属性、反函数属性、传递属性和对称属性。

(6) 为类添加属性约束 (Restrictions)。

在 OWL 本体中, 属性可以用于构建属性约束, 从而对类进行定义和描述。属性约束定义了一个新的匿名类, 该类中的所有个体均满足限制, 所描述的类是这个匿名类的子类, 譬如所有具有至少一个题名的个体构成了一个匿名类, `co:Document` 类是这个类的子类。OWL 本体具有三种属性约束: 量词约束、基数约束和取值约束。详细关于 `skos:Concept` 类的子类基于属性 `skos:inScheme` 的取值约束见表 8.8。

表 8.8 核心元数据本体中 `skos:Concept` 类及其子类的取值约束

skos:Concept 类及其子类	取值的个体
skos:Concept	所有采用 SKOS 语言表示的受控词表
skos:LanguageConcept	<ul style="list-style-type: none"> • http://www.nlc.gov.cn/vocab/Lang_v1.0 • http://www.nlc.gov.cn/vocab/ISO639-2 • http://www.nlc.gov.cn/vocab/RFC4646 • http://www.nlc.gov.cn/vocab/EthnicLang
skos:ResourceTypeConcept	<ul style="list-style-type: none"> • http://www.nlc.gov.cn/vocab/DCMIType • http://www.nlc.gov.cn/vocab/IRType
skos:MediaFormatConcept	<ul style="list-style-type: none"> • http://www.nlc.gov.cn/vocab/IMT.rdf
skos:LocationConcept	<ul style="list-style-type: none"> • http://www.nlc.gov.cn/vocab/CCT_CLC_Place_v4.0 • http://www.nlc.gov.cn/vocab/TGN • http://www.nlc.gov.cn/vocab/TGS • http://www.nlc.gov.cn/vocab/ISO3166

(7) 为类和属性添加注释属性 (Annotation Properties)。

本体中除了数据属性和对象属性之外, 还有另外一种属性, 称为注释属性。注释属性的作用是对本体中的类、属性和个体 (即实例) 进行注释说明。OWL 本体中有五个预定义的注释属性可以直接使用, 对本体中的类/属性/个体提供注释信息。

核心元数据本体中 `co:Document` 类的注释属性如表 8.9 所示, `dc:title` 属性的注释属性如表 8.9 所示。

表 8.9 核心元数据本体中 co:DocumentResource 类的注释属性

owl:versionInfo	1.0
rdfs:label	文献信息资源（简称文献资源）
co:definition	文献信息资源是图书馆中各种纸质文献资源、数字化文献资源和网络信息资源的统称
rdfs:comment	包括各类型和网络信息资源
rdfs:seeAlso	
rdfs:isDefinedBy	http://www.nlc.gov.cn/onto/core_v1.0.owl

（8）创建一个实例。

现给出《数字图书馆的知识组织系统：从理论到实践》一书的 DC 元数据（表 8.10），图 8.5 表示该书元数据的简单语义化表示（RDF/XML 格式）以及图 8.6 展示其 RDF 图。

表 8.10 图书《数字图书馆的知识组织系统：从理论到实践》基于 DC 的元数据

元数据元素	元素中文标签	元 素 值
dc:title	题名	数字图书馆知识组织:从理论到实践
dc:creator	创建者	王军
dc:publisher	出版者	北京大学出版社
dc:language	语种	chi
dc:subject	主题	数字图书馆
dc:identifier	标识符	9787301149034 (ISBN 号)
dc:type	资源类型	Text（采用 DCMI 类型词表中的术语）
dc:format	格式	200 页; 23cm

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:dc=http://purl.org/dc/elements/1.1/
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  <rdf:Description rdf:about="http://www.nlc.gov.cn/resource/004106310">
    <dc:title xml:lang="zh">数字图书馆知识组织:从理论到实践</dc:title>
    <dc:creator xml:lang="zh">王军</dc:creator>
    <dc:publisher xml:lang="zh">北京大学出版社</dc:publisher>
    <dc:language xml:lang="zh">chi</dc:language>
    <dc:subject xml:lang="zh">数字图书馆</dc:subject>
    <dc:identifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">9787301149034</dc:identifier>
    <dc:type xml:lang="en">Text</dc:type>
    <dc:format xml:lang="zh">200 页; 23cm</dc:format>
  </rdf:Description>
</rdf:RDF>
```

图 8.5 图书《数字图书馆的知识组织系统：从理论到实践》元数据的简单语义化表示（RDF/XML 格式）

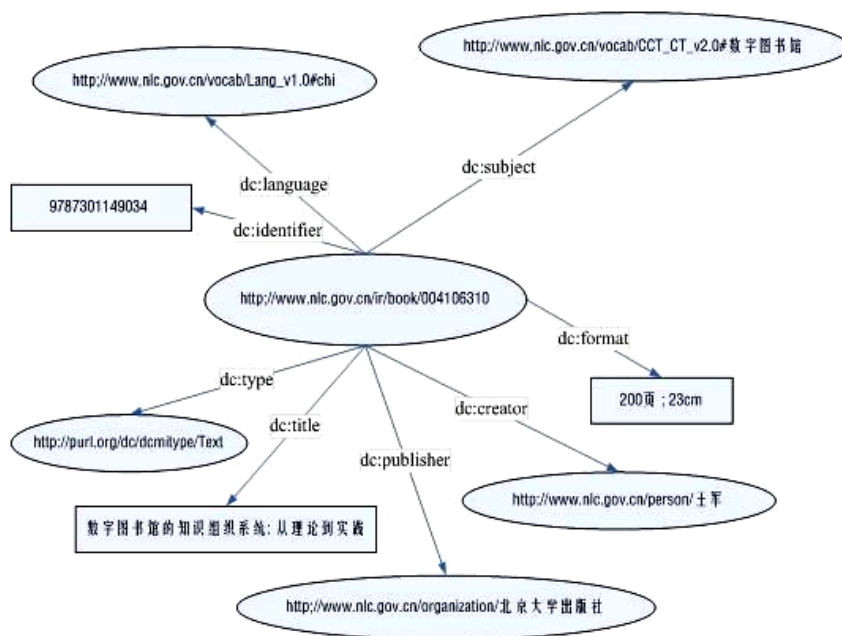


图 8.6 图书《数字图书馆的知识组织系统: 从理论到实践》元数据的 RDF 图

8.3.6 本体描述语言

1) RDFS

RDFS^① (Resource Description Framework Schema, 资源描述框架模式) 是 W3C 在 RDF 的基础上制定的 RDF 词汇描述语言标准, 主要是扩充了 RDF 对于资源描述的能力。目前 RDFS 已经有相关的 W3C 推荐标准^②, 结合 RDF 一起有着广泛的研究和应用。

RDFS 使用一种机器可以理解的体系来定义描述资源的词汇, 其目的是提供词汇嵌入的机制或框架, 在该框架下多种词汇可以集成在一起实现对 Web 资源的描述。RDFS 引入一个面向对象、可扩展的类型 (Type) 系统到 RDF 中提供方法定义合适的定义域和值域以及类和子类层次。在 RDFS 中, 属性不是局部于一个类 (如面向对象语言中的属性), 而是全局地在属性连接的各类中描述。模式文档的作用类似一个合同作用: 在提交一个 RDFS 的同时应用程序开发者保证他们的应用程序可与依从 RDFS 的 RDF 实例数据工作。RDFS 说明书提供了核心类、核心属性类型和核心限制, 三个核心类由 RDFS 机构提供: 资源 (即所有对象的类), 属性类型 (即所有二元关系的类), 类 (即所有类型的类)。有两个核心属性类型: instanceOf (定义一个资源与类的一个元素间的关系) 和 subClassOf (定义类的两个元素间的关系)。限制有两个核心实例: Range 和 Domain 分别定义属性类型的值域和定义域。

2) OIL, DAML 及 DAML+OIL

OIL^③ (OIL, Ontology Interchange Language/Ontology Inference Layer, 本体互换语言/本体推理层) 由美国斯坦福大学、荷兰阿姆斯特丹大学、曼彻斯特大学计算机科学系、贝尔实

① RDF[EB/OL].[2013-06-10].<http://www.w3.org/RDF/>.

② RDF Schema 1.1[EB/OL].[2013-06-10].<http://www.w3.org/TR/rdf-schema/>.

③ OIL[EB/OL].[2013-06-10].<http://www.ontoknowledge.org/oil>.

验室、AIFB、麻省理工学院等多家机构从 2000 年起联合开发,以项目 Ontoknowledge 为依托。OIL 的主要功能分为两类:合并和表示本体及进行系统间交互,其设计目的是提供尽可能多的建模元语,可用于基于框架系统和 DL 的本体。但目前 OIL 已经被 OWL 所取代,不再使用了。

OIL 的设计借鉴了三个方面:由 DL 提供正规语义和有效推理,由框架系统提供丰富的建模元语,由 XML 和 RDF (S) 提供通用形式化语法。(事实上,OIL 有着 XML 和 RDFS 两种不同的语法体系。)在 OIL 中,本体是由一系列元素和组件组成的,并被划分为三个层次:对象层(Object Level)、本体定义层(Ontology Definition)及本体容器层(Ontology Container)。另外,OIL 本身是分层次的,每一个层次都在前一个层次基础上添加了功能性(Functionality)和复杂性(Complexity)。

DAML^①(DARPA Agent Markup Language, DARPA 标记语言)是由 DAML 委员会主持,从 2000 年开始开发设计的一种本体表示语言。DAML 完全是以 XML 和 RDF (S) 为基础的,其目的是通过扩展 RDF (S) 的语义表示能力从而形成新的本体表示语言。DAML 最初版本为 DAML0.5,后来发展为 DAML-ONT,在 2000 年年末,DAML 与 OIL 相结合,形成新的语言标准——DAML+OIL,因而,DAML+OIL 可视为 DAML 的最新版本。目前 DAML+OIL 也已经被 OWL 所取代,不再使用。

DAML-ONT 是建立在 XML 和 RDF (S) 基础上的,引入了许多实用的新特性。但 DAML-ONT 缺乏对 DL、框架系统的借鉴,因而 DAML-ONT 作为一种本体表示语言,在一定程度上与 RDFS 表示本体有着类似的问题。但 DAML-OIL 及时地借鉴了 OIL 开发的成果,在 DAML-ONT 基础上引入了许多 OIL 的语言组建,并提供对 DL 的有力支持,从而形成了较为完善的 DAML-OIL 语言。

OIL、DAML、DAML+OIL 虽然没能成为最终的推荐标准,并最终不再使用,但需要看到的是它们对于 NKOS 表示语言,尤其是网络本体表示语言的重要贡献,这主要表现在:第一,它们是第一批建立在 XML、RDF (S) 标准之上的网络本体表示语言,这与先前的低级网络本体表示语言(如 SHOE、XOL 等)有着本质的不同,它们体现了现行网络构架向 SW 构架发展的努力;第二,它们第一次将 DL 引入网络本体表示,并将其作为极为重要的内在逻辑形式,使得 DL 在 NKOS 中得到广泛应用,而将本体语言与现存逻辑形式相结合,成为赋予本体语言正规语义和推理支持的一种有效手段;第三,它们的开发为后来 OWL 的开发积累了宝贵的经验(许多 OIL、DAML 的开发人员进一步参与了 OWL 的开发),提供了坚实的基础,并且为后来 OWL 得到广泛的认同和应用做出了贡献。

3) OWL 1.0

OWL^②(Web Ontology Language, 网络本体语言)是 W3C 在 2004 年制定的一种语言标准,主要用来表示网络本体(Web Ontology)这种 NKOS,目前已经形成多篇规范文档,作为语义网语言栈的重要组成部分受到广泛关注。

1999 年提出 Semantic Web 的设想之后,陆续出现了 OIL、DAML 等多种网络本体表示语言,作为互联网最权威的标准化组织 W3C 也开始制定一种语言来作为推荐标准,在这样的背景下 OWL 诞生了。OWL 一方面吸取了 OIL、DAML 等的开发经验,另一方面也考虑到 SW 的构架对本体语言的要求。因而,OWL 一方面吸取了成功嫁接在 OIL 上的 DL 作为内在逻辑模式;另一方面 OWL 和 DAML 一样,是建立在 RDF(S)的基础上的,另外,OWL 还借鉴了 OIL 的分层次语言思想,提供了三个不同功能的子语言。OWL 出现后,各方面的评价

① The DARPA Agent Markup Language Homepage[EB/OL].[2013-06-10]. <http://www.daml.org>.

② OWL[EB/OL].[2013-06-10]. <http://www.w3.org/owl>.

基本上是积极的;又由于是 W3C 的推荐标准, OWL 得到了广泛的讨论、研究和应用尝试,在其基础上的开发工具也纷纷面世。

(1) OWL 层次。

OWL 包含三个子语言: OWL Full、OWL DL、OWL Lite。OWL Full 是完整的 OWL 语言,拥有最丰富的建模元语和最强大的描述能力,支持那些需要最强的表达能力和完全自由 RDF 语法的用户。OWL Full 与 RDF (S) 是完全兼容的:任何合法的 RDF 文档都是合法的 OWL 文档,任何有效的 RDFS 推论都是有效的 OWL 推论。而 OWL Full 过强的表示能力带来的问题是 OWL Full 是不可判定的,到目前为止还没有任何软件能够完全支持 OWL Full 的推理。

OWL DL 是完全基于 DL 的,为此它对 OWL Full 中建模元语的使用方式施加了约束:最主要的是禁止了 OWL Full 中资源的多重身份,从而使得 OWL DL 的计算是完备而可判定的,保障了有效推理的可能性。OWL DL 用于支持那些既需要强表达能力也需要保证计算完备性 (Computational Completeness, 即所有的结论都能够确保被计算出来) 和可判定性 (Decidability, 即所有的计算都能在有限的时间内完成) 的用户。不利的一面是 OWL DL 放弃了与 RDF(S) 的完全兼容:虽然合法的 OWL 文档仍然是合法的 RDF 文档,但合法的 RDF 文档可能需要经过一系列处理才能生成合法的 OWL 文档。

OWL Lite 是对 OWL DL 的进一步限定,取消了一些不常用的建模元语,虽然这会导致 OWL Lite 的表示能力比 OWL DL 更低,但这使得用户更容易掌握这种语言,在其基础上的开发工具也更容易设计与使用。OWL Lite 用于提供给那些只需要一个分类层次和简单约束的用户。

使用 OWL 的本体开发者要考虑哪个子语言最符合他的需求。选择 OWL Lite 还是 OWL DL 主要取决于用户在多大程度上需要 OWL DL 提供的更强的表达能力;选择 OWL DL 还是 OWL Full 主要取决于用户在多大程度上需要 RDF Schema 的元建模 (Meta-modeling) 机制,以及是否需要高效推理。

(2) OWL Full 建模元语。

在最新的 OWL 语义规范标准中^①,总计包括 51 个建模元语,其中 11 个来自于 RDF (S),另外 40 个为 OWL 所独有的建模元语。根据其作用不同,下文将其分 8 个大类,选择较为主要的建模元语在下文进行介绍。

① 首部相关。

由于 OWL 文档都是 RDF (S) 文档,因而 OWL 文档的根元素同样是 rdf:RDF 元素。在 OWL 文档的首部,有一些声明该文档的管理信息的代码,它们都包含在一个 owl:Ontology 元素之下,具体包括注释 (rdfs:comment)、标签 (rdfs:label)、版本信息 (owl:priorVersion 等)、对其他本体的引用 (owl:imports)。其中需要说明的是 owl:imports 元语,它声明了该本体引用了哪些其他本体,是一个具有传递性的属性。

② 类。

OWL 中定义类的元语为 owl:Class,它是 rdfs:Class 的下位类,其使用方法也与 rdfs:Class 一致。OWL 中元语定义类间关系不仅限于下位类,还可以通过元语“不相交类”(owl:disjointWith)来声明两个类是不相交的,以及通过元语“等同类”(owl:equivalentClass)来声明两个类是等同的。另外,OWL 预先定义了两个类:owl:Thing 和 owl:Nothing。owl:Thing 是所有实例的类,是所有 OWL 类的上位类;owl:Nothing 则没有任何实例,是任意 OWL 类

① OWL Web Ontology Language Semantics and Abstract Syntax[EB/OL].(2009-11-12) [2013-06-10].
<http://www.w3.org/TR/owl-semantics>.

的子类。例如，我们可以声明网站（Website）和链接（Link）都是网络资源（NetResource），但它们是不相交的类。

```
< owl:Class rdf:ID = "Website" >
  < rdfs:subClassOf rdf:resource = "#NetResource" />
  < owl:disjointWith rdf:resource = "#Link" />
< /owl:Class >
```

③ 属性。

在 OWL 中存在两种属性：对象属性（owl:ObjectProperty）和数据属性（owl:DatatypeProperty），它们都是 rdf:Property 的子类。前者的定义域和值域都是“目标”，即 RDF 中的“资源”。后者的值域是某种类型的数值，即 RDF 中的“常量”，如上例中“Creator”是目标属性，而“Email”是数据属性。OWL 允许用元语 owl:inverseOf 来声明逆属性。属性 B 是属性 A 的逆属性，即属性 A 的定义域是属性 B 的值域，属性 B 的定义域是属性 A 的值域，且 A 与 B 在语义上是互逆的，如“Creator”和“isCreatedBy”即为一对互逆属性。属性间等同的关系则通过元语 owl:equivalentProperty 来声明，如可以认为“Creator”和“Writer”是等同的属性。例如，我们可以声明“Creator”是一种对象属性，它有一个逆属性是“isCreatedBy”，并且和属性“Writer”是等同的。

```
< owl:ObjectProperty rdf:ID = "Creator" >
  < owl:inverseOf rdf:resource = "#isCreatedBy" />
  < owl:equivalentProperty rdf:resource = "#Writer" />
< /owl:ObjectProperty >
```

④ 属性约束。

属性约束是 OWL 非常强大的表示机制，可以用于多个方面的描述。其原理是：一个类 A 是另一个类 B 的子集，如果类 A 所有的元素都符合一定条件，而所有符合这些条件的元素构成了类 B，整个过程中可以不对类 B 进行声明——类 B 可以是匿名的。属性约束有两种方式：一种是值域约束，相关元语包括 owl:allValuesFrom、owl:hasValue、owl:someValuesFrom；另一种是基数约束（Cardinality Restriction），相关元语包括 owl:minCardinality 和 owl:maxCardinality。不论是哪种约束，都是在元素 owl:Restriction 中通过 owl:onProperty 声明对哪个属性进行约束的。下面举例说明：定义学术网站（AcademicWebsite）这种资源，它是网站的子类，并且认为的创建者只能由学术人员（AcademicStaff）担任，那么可以表示如下。

```
< owl:Class rdf:ID = "#AcademicWebsite" >
  < rdfs:subClassOf rdf:resource = "#Website" />
  < rdfs:subClassOf >
    < owl:Restriction >
      < owl:onProperty rdf:resource = "#isCreatedBy" />
      < owl:allValuesFrom rdf:resource = "#AcademicStaff" />
    < /owl:Restriction >
  < /rdfs:subClassOf >
< /owl:Class >
```

如果认为学术网站的创建者只能由某个具体的人担任，那么只需要将上面代码中的“owl:allValuesFrom”替换成“hasValue”，并且指向某个具体资源即可。如果认为学术网站的创建者当中必须包含学术人员，而不必全都是学术人员，则只需要将上面代码中的“owl:allValuesFrom”替换成“owl:someValuesFrom”即可。

如果认为一个学术网站的创建者最少有 1 个，最多不超过 10 个，那么可以使用基数限制。

```
< owl:Class rdf:about = "#AcademicWebsite" >
  < rdfs:subClassOf >
    < owl:Restriction >
      < owl:onProperty rdf:resource = "#Create" />
      < owl:minCardinality
        rdf:datatype = "&xsd;nonNegativeInteger" >
        1
      < /owl:minCardinality >
    < /owl:Restriction >
  < /rdfs:subClassOf >
  < rdfs:subClassOf >
    < owl:Restriction >
      < owl:onProperty rdf:resource = "#Create" />
      < owl:maxCardinality rdf:datatype = "&xsd;nonNegativeInteger" >
        10
      < /owl:maxCardinality >
    < /owl:Restriction >
  < /rdfs:subClassOf >
< /owl:Class >
```

⑤ 属性特性。

这些属性类建模元语用来声明属性及其值的相关特性。具体而言，包括如下元语：传递属性（owl:TransitiveProperty），用来声明一个属性是传递性的，如“更大”、“更高”等；对称属性（owl:SymmetricProperty），用来声明一个属性是对称的，如“是朋友关系”，“有一样的成绩”等；函数属性（owl:FunctionalProperty），用来声明一个属性是函数性质的，即对于每一个实例，该属性的属性值只有一个，如“年龄是”、“身份证号码是”等；反函数属性（owl:InverseFunctionalProperty），用来声明一个属性是反函数性的，即该属性的逆属性是函数属性的，或者说，给定该属性值域中的一个值，定义域中有且只有一个实例与之对应，如“个人电话号码是”、“身份证号码是”等。

⑥ 布尔组合。

这些属性类建模元语建立了类和类之间的布尔组合操作，可以在原有定义的类的基础上通过布尔组合生成新的类，这些类可以是匿名的。具体而言，包括如下元语：补（owl:complementOf），用来声明一个类的补类，可以用来声明 A 与 B 不相交，即 A 是 B 的补类的子类；并（owl:unionOf），用来声明一个类是由多个类的合并部分组成的；交（owl:intersectionOf），用来声明一个类是由两个或多个类相交的相交部分组成的；例如，我们可以通过声明网站是链接的补类的子类，从而声明网站和链接是不相交的类。

```
< owl:Class rdf:about = "#Website" >
  < rdfs:subClassOf >
    < owl:Class >
      < owl:complementOf rdf:resource = "#Link" />
    < /owl:Class >
  < /rdfs:subClassOf >
```

```
</owl:Class>
```

可以定义网站资源（WebsiteResource）这个类，它是网站和链接的并集，代码如下。

```
< owl:Class rdf:ID = "WebsiteResource" >
  < owl:unionOf rdf:parseType = "Collection" >
    < owl:Class rdf:resource = "#Website"/>
    < owl:Class rdf:resource = "#Link"/>
  </owl:unionOf>
</owl:Class>
```

⑦ 其他。

列举：OWL 中允许通过列举一个类的所有实例的方式来定义一个类，这个功能是通过元语 owl:oneOf 来实现的。

版本信息：版本信息一般用于本体的管理，相关的原语有：owl:priorVersion、owl:versionInfo、owl:backwardCompatibleWith、owl:incompatibleWith 等。这些属性对建模没有太大的作用。

数据类型：OWL 中可以声明常量的数据类型，如可以声明数值属性“姓名”的值为字符串型常量（xsd:string），数值属性“年龄”的值为正整数型常量（xsd:integer）。OWL 没有定义关于数据类型的建模元语，而是直接引用 XMLS 中的数据类型作为 OWL 的建模元语，这体现了层次化语言的优越性。

(3) OWL 推理。

一般而言，语言的表示能力越强，在其基础上的有效推理就越困难。OWL Full 有着最为丰富的表示能力，但目前 OWL Full 的有效推理还没有得到保证。另一方面，OWL Lite 的推理效率是最高的，但它的表示能力受到很大限制。OWL DL 是两者的折中，有着良好的表示能力和推理性能，又由于 OWL DL 完全是以合法的 DL 为内在逻辑模式的，因而这里着重介绍 OWL DL 的推理。

一种知识表示语言的推理能力往往并非这种语言本身定义的，而是通过将该语言映射到已有的逻辑模式上，借用已有逻辑模式的推理能力来作为该语言自身的推理能力。OWL DL 正是通过这种方式获取了正规化的语义和可保障的推理能力。OWL DL 通过对 OWL Full 的一些削减和约束，使得其自身能够恰当地映射到 DL-SHIQ 上。具体而言，OWL DL 中所有的建模元语考虑了以下 8 个方面：类(Class)、数据类型(Datatype)、对象属性(Object Property)、数值属性(Datatype Property)、个体(Individual)、数据值(Data Value)、本体属性(Ontology Property)及注释属性(Annotation Property)。其中后两者是对一些对象的说明，提供给使用者以方便使用或维护等，而并没有对领域信息进行描述，因而在映射过程中不予考虑。而前六者则分别与 DL-SHIQ 中的概念(Concept)、具体域(Concrete Domain)、角色(Role)、特征(Feature)、个体(Individual)和值(Value in Concrete Domain)相对应。在此基础上，OWL DL 中所有建模元语及其使用方式都可以映射到 DL-SHIQ 中，从而使整个 OWL DL 语言被映射到 DL-SHIQ 上，可以借助建立在 DL 基础上成熟的推理机制，如 FACT、RACER、DLP 等完成 OWL DL 的推理工作。

总的来说，OWL 是建立在 RDF(S) 基础之上的，与后者相比，OWL 的性能主要在两个方面得到了提升：第一，OWL 提供了更丰富的建模元语和组合机制，使得 OWL 的表示能力更强；第二，OWL 以 DL 为内在逻辑模式，增强了推理能力。

作为一种网络本体表示语言，OWL 在设计过程中借鉴了多种本体语言的开发经验，考

虑到多方面的需求，它的三个子语言在描述能力和推理能力之间提供了三种不同的选择，极大地扩展了其使用空间。OWL 一方面建立在 RDF(S)的基础上，使得其作为 SW 的层次语言能够很好地向下兼容；另一方面 OWL 以 DL 为内在逻辑模式，不但可以借用 DL 的推理能力，还保障了语言的一致性。最后，OWL 作为 W3C 的推荐标准，统一了网络本体的表示语言，使得不同组织制定的本体能够相互利用，促进了知识共享，充分体现了 SW 的设计初衷。最后，OWL 是基础标准，有着良好的扩展性，目前在其基础上已经出现了 OWL-S 语义网服务标识语言、OWL-QL 语义网查询语言等扩展语言。

另一方面，OWL 并非十全十美。首先，OWL 采用 DL 为内在逻辑模式，而 DL 的性能是否适合于网络知识表示还有待时间检验——事实上，单论推理能力，OWL 并没有 Loom、FLogic 等非网络知识表示语言强大，它还需要 Jena 等后期开发工具的大力支持。另一个问题涉及本体在网络应用中自身的尴尬：本体作为一种对共享知识体系的描述，其本身具有静态性、稳定性、认同性。而网络知识是快速变化的、不稳定的，而且许多网络知识并没有得到大部分的一致认同。这就给本体在网络环境下的应用带来了较大的困难。OWL 并没有提供相应的动态机制，由此描述的本体缺乏自动和半自动的动态更新能力，这是 OWL 在网络上得到大规模应用所必须解决的问题。

4) OWL 2.0

OWL 不是语义的终结，语义的复杂化使得不断有新的标准推出。2009 年 10 月 W3C 推出了 OWL 的新版本 2.0。OWL 2.0 与原有的 OWL（即 OWL 1.0）兼容，但添加了一些新的特征（详见表 8.11），如对属性增强的表达能力、对数据类型的扩展支持、简单的元模型能力、扩展的注释能力等，具体表现在以下五个方面：

表 8.11 OWL2.0 的新特性^①

类表达式	本地自反（Local Reflexivity）（自我限制） 限定精确/最大/最小基数限制的对象和数据 在 n 元数据值域上的全称命题（Universal）和存在命题（Existential）限制
类公理 （Class Axioms）	两两不相交类 类不相交并
属性表达式 （Property Expressions）	通用对象属性和空对象属性 通用数据属性和空数据属性 逆对象属性表达式
属性公理 （Property Axioms）	属性链包含 不相交对象属性 不相交数据属性 自反，非自反和非对称对象属性
数据值域 （Data Ranges）	数据类型定义 数据值域补集，交集和并集 数据类型限制和分面（Facets） n 元数据类型的钩连（Hook）
断言 （Assertions）	否定的对象属性断言 否定的数据属性断言

① 曾新红,蔡庆河. OWL2 Web 本体语言快速参考指[EB/OL].(2012-09-25)[2013-06-10].<http://nkos.lib.szu.edu.cn/OWL2/OWL2QuickReferenceGuideSimplifiedChinese.htm>

续表

注释 (Annotation)	注释断言 公理或注释的注释 注释子属性 注释属性定义域和值域 owl:deprecated 注释属性
附加内置数据类型 (Extra Built-in Datatypes)	owl:rational, owl:real, xsd:dateTimeStamp, rdf:PlainLiteral
其他 (Others)	键 声明 元建模能力 (Metamodeling Capabilities) (双关 (Punning)) 匿名个体

(1) 语法使得很多常见的陈述能够更容易的构建, 这使得 OWL 语言更容易使用;

(2) 新功能提高了表达能力, 包括一些新属性, 比如 Reflexive, Irreflexive 和 Asymmetric 等, 同时一些新的基数限制也极大的增强了该语言的表达能力, 还有一些新的功能, 如 Property Chains 和 Keys;

(3) 其增加了对数据类型的支持, 包括更多由 OWL2.0 提供的内置数据类型, 同时, OWL2.0 也允许用户在创建 ontologies 的时候, 自己定义数据类型;

(4) 简单的元模型能力和扩展的注释能力。元模型能力包括一个新功能叫 Punning。注释在 OWL2.0 中很强大, 用户可以给公理添加注释、可以给注释属性添加定义域和值域信息以及给注释自身添加注释;

(5) 新的子语言: 包括 OWL 2 EL、OWL 2 QL 和 OWL 2 PL, 这些语言可以在表达性和有效性中找到权衡, 给用户提供更多选择。

8.3.7 基于本体的信息组织实例

本节通过一个实例来说明如何将本体应用于信息组织。以《信息资源管理理论》一书为例, 将该书来自图书馆书目数据库的 MACR 元数据和来自万方数据库的 NoteFirst^①两种元数据转化为统一的 RDF 格式的元数据来实现不同类型元数据的融合。

该书的 MACR 元数据显示如下:

```
000 01045nam 2200325 450 (记录头标区)
001 0000568854
005 20031225163800.0
101 0_ |a chi (作品种类为中文)
200 1_ |a 信息资源管理概论 |A Xin Xi Zi Yuan Guan Li Gai Lun |f 主编孙建军 (题名与责任说明项: 题名、拼音及责任者)
210 __ |a 南京 |c 东南大学出版社 |d 2003
(出版发行项: 出版地、出版社、出版时间)
215 __ |a 303 页 |c 图 |d 26cm (载体形态项: 页数、其他形态、尺寸)
701 _0 |a 孙建军 |A Sun Jian Jun |4 主编 (责任者, 责任方式)
```

① NoteFirst 是一个专属于科研技术人员的信息获取, 文献管理以及知识共享的服务平台; 也是科技文献服务市场上第一款把科技文献管理和开放存取、科技资源交流共享相结合的服务系统。

该书的 NoteFirst 的元数据显示如下：

```
<?xml version="1.0" encoding="utf-8"?>
<Bibliographies>
  <BibliographiesCount>1</BibliographiesCount>
  <Bibliography>
    <Type>Book</Type>
    <PrimaryTitle>
      <Title>信息资源管理概论</Title>
    </PrimaryTitle>
    <Authors>
      <Author>
        <Info>
          <FullName>孙建军</FullName>
        </Info>
      </Author>
    </Authors>
    <PrintDate>2003</PrintDate>
    <PageScope>312</PageScope>
    <Publisher>东南大学出版社</Publisher>
    <Language>zh-CHS</Language>
  </Bibliography>
</Bibliographies>
```

针对同一本书的两种不同格式的元数据，基于元数据本体，可将这种元数据都转换成相同的 RDF 格式表示，显示如下：

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:metaonto="http://example.nju.edu.cn/Ontology/metaonto.owl#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
  <metaonto:Book rdf:about="http://example.nju.edu.cn/book/7810893211">
    <dcterms:title xml:lang="zh">信息资源管理概论</dcterms:title>
    <dcterms:creator rdf:resource="http://example.nju.edu.cn/person/Sunjianjun"/>
    <dcterms:publisher
      rdf:resource="http://example.nju.edu.cn/organization/SoutheastUniversityPress"/>
    <dcterms:date
      rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">2003</dcterms:date>
    <dcterms:format xml:lang="zh">303 页</dcterms:format>
    <rdfs:isDefinedBy rdf:resource="http://example.nju.edu.cn/book/7810893211.rdf"/>
  </metaonto:Book>
</rdf:RDF>
```

8.4 基于语义网的网络知识组织系统

8.4.1 SKOS 语言简介

SKOS^① (Simple Knowledge Organization System, 简单知识组织系统) 是 W3C 于 2005 年制定的一个语言规范, 主要用于描述叙词表、分类法、主题词表等受控词表。

在目前实际应用的 NKOS 中, 基于传统知识组织系统的、非本体的初级工具占了主要部分。以受控词表为例, 由于长时间以来缺乏相关的规范标准, 网络上出现了多种受控词表表示标准——一方面各种词表标准相互竞争, 取长补短, 各自完善; 另一方面这种多种标准共存的局面显然不利于不同 NKOS 间的互操作, 有碍于知识共享。SKOS 的出现旨在解决这个问题, 一方面, SKOS 结构简单易用, 借鉴了许多现有知识组织系统的标准 (主要是词表和分类法); 另一方面, 它的表示能力又比较强, 可以用于从受控词表到分类法等多种 NKOS 的表示。

实际上, SKOS 可以视为 RDF(S) 和 OWL 在 NKOS 表示领域的特定应用, SKOS 中的建模元语本身就被定义为 OWL 中的类或者属性, 它们之间的关系也可以使用 OWL 中的关系进行描述, 因而可以将 SKOS 规范本身视为可以使用 OWL 表示的本体。这样, SKOS 一方面丰富了 RDF(S) 略有不足的表示能力, 并继承了 OWL 的部分推理能力, 另一方面又避免了完全使用 OWL 系列语言带来的复杂性, 不失为一个较好的折中方案。

使用 SKOS 表示 NKOS 的原理非常简单: 对于 NKOS 的每一个概念, 都视为 RDF 图中的一个节点, 而概念与概念之间、概念与常量之间的关系则被视为 RDF 图中的有向边, 将概念和常量连接在一起。每个受控词表都可以按照上述方式表示成一个 RDF 有向图, 这个图本身可以用 RDF/XML 进行描述。作为一种语言, SKOS 的任务是提供一系列词汇及其语法, 这些词汇对应着 NKOS 中的概念和关系, 这样用户就可以使用 RDF 图这种统一的方式来记录 NKOS, 从而使 NKOS 发挥更大的作用。下面就介绍 SKOS 的建模元语。

1) SKOS 建模元语

在最新的 SKOS 参考规范中^②, 共定义了 32 个 SKOS 建模元语, 它们按照不同作用被划分成了 8 个大类, 下面简要介绍 SKOS 的建模元语。

① 概念。

概念 (skos:Concept) 是 SKOS 最基本的建模元语, 它被定义为一个 OWL 类, 用来声明或定义某个资源是一个概念性 (Conceptual) 的资源, 即 NKOS 中的一个“概念”, 如“爱”这个概念, 见图 8.7。

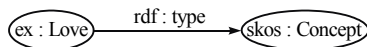


图 8.7 skos:Concept 元语的使用: 一个例子^③

相应地用 XML/RDF(S) 语法表示如下。

```
<rdf:RDF
```

```
  xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
```

① SKOS Simple Knowledge Organization System-Home Page[EB/OL].[2013-06-10].<http://www.w3.org/2004/02/skos/>.

② SKOS Simple Knowledge Organization System Reference[EB/OL].(2008-01-25)[2013-06-10].
<http://www.w3.org/TR/2008/WD-skos-reference-20080125/>

③ prefix ex:<<http://www.example.com/concepts#>>, 下同, 这里只是一个形式化的例子, 没有实际意义。

```

< rdf:Description
  rdf:about = "http://www.example.com/concepts#love">
  <rdf:type rdf:resource =
    "http://www.w3.org/2004/02/skos/core#Concept"/>
< /rdf:Description >
< /rdf:RDF >

```

如果某个资源被定义为概念，那么在 SKOS 中就可以用 `skos:Concept` 元语以如下语句直接声明它是一个概念，而不必再使用 RDF 中的 `rdf:Description`、`rdf:type` 元语来重复定义了^①。

```

.....
< skos:Concept rdf:about = "#ID" >
.....
< /skos:Concept >
.....

```

② 概念体系。

通常情况下 SKOS 中的概念并非是孤立的，而是和其他概念相联系的，共同形成一个体系。一个概念体系就是指一系列概念及其间语义关系的声明。SKOS 中的概念体系（`skos:ConceptSchema`）就是用来声明和定义某个资源是一个概念体系的，它被定义为一个类。与此相关的是两个属性：属于框架（`skos:inSchema`）和包含顶级概念（`skos:hasTopConcept`）。前者用来声明某个概念属于某个概念体系；后者用来声明某个概念体系中，一个概念链的最顶层概念是什么，它们都被定义为 OWL 对象属性。见图 8.8。

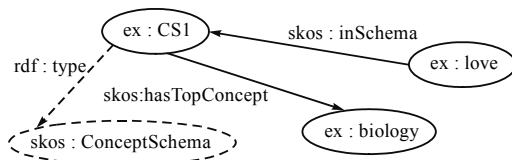


图 8.8 SKOS 概念框架相关元语的使用：一个例子

相应地用 RDF/XML 语法表示如下。

```

"< rdf:RDF
  xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos = "http://www.w3.org/2004/02/skos/core#" >
  < skos:ConceptSchema rdf:about =
    "http://www.example.com/conceptschema#CS1" >
    < skos:hasTopConcept rdf:about =
      "http://www.example.com/concept#biology"/>
  < /skos:ConceptSchema >
  < skos:Concept rdf:about =
    "http://www.example.com/concept#love" >
    < skos:inSchema rdf:about =

```

① 下面要介绍的元语 `skos:ConceptSchema` 的使用方法与此相类似。


```

    "http://www.example.com/conceptschema#CS1"/>
  </skos:Concept>
</rdf:RDF>

```

③ 词汇标签。

这些属性为资源添加某种记号，具体包括以下几个属性：首选标签（`skos:preLabel`）、可选标签（`skos:altLabel`）和隐藏标签（`skos:hiddenLabel`），它们都被定义为 OWL 数据属性，它们之间两两不相交。首选标签是资源在给定语言下的首选的词汇标签，在同一个给定了语言的受控词表中，两个概念的首选标签的值是不能相同的；一个资源也不能有两个不同的（同语言的）首选标签。可选标签是资源在给定语言下的可以选用的词汇标签，一般情况下首选标签和可选标签的值应该是同义字，但对于概念则不必做如此严格的要求。隐藏标签是资源的一种标签，这种标签在资源可视化展示的时候应该是隐藏的，而在进行自由文本检索的时候却是可以使用的。这种标签最常用的就是标记某个词汇的常见错误拼写。具体示例见图 8.9。

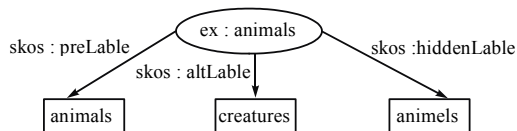


图 8.9 SKOS 标签属性的使用：一个例子

相应地用 XML/RDF(S)语法表示如下。

```

<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#">
  <skos:Concept
    rdf:about="http://www.example.com/concepts#animals">
    <skos:prefLabel> animals </skos:prefLabel>
    <skos:altLabel> creatures </skos:altLabel>
    <skos:hiddenLabel> animels </skos:hiddenLabel>
  </skos:Concept>
</rdf:RDF>

```

④ 注释属性。

注释属性为资源提供某些相关注释，包括如下建模元语：注释（`skos:note`）、定义（`skos:definition`）、范围注释（`skos:scopeNote`）、范例（`skos:example`）、历史注释（`skos:historyNote`）、编辑注释（`skos:editorialNote`）和变更注释（`skos:changeNote`），它们都被定义为 OWL 的对象属性，其中 `skos:note` 是其他 6 个属性的父属性。

W3C 的规范文档对于这些属性的使用给出了三种推荐方式，事实上是注释属性的值域的三种情况：在第一种情况下值域是 XML 文本字符串；在第二种情况下值域是对某个资源的一系列说明；在第三种情况下值域是 URI。这三种情况下 SKOS 的语法和 RDF 语法完全一致，只不过 RDF 中的具体谓词在这里以建模元语的形式出现。

⑤ 语义关系。

这些属性用于声明概念间的语义关系（图 8.10），具体包括如下建模元语：语义关系（`skos:semanticRelation`）、相关关系（`skos:related`）、上位类传递关系（`skos:broaderTransitive`）、上位类关系（`skos:broader`）、下位类传递关系（`skos:narrowerTransitive`）及下位类关系

(skos:narrower), 它们都被定义为 OWL 对象属性, 其中 skos:related、skos:broaderTransitive 和 skos:narrowerTransitive 都是 skos:semanticRelation 的子属性; 而 skos:broader 和 skos:narrower 又分别是 skos:broaderTransitive 和 skos:narrowerTransitive 的子属性。另外, skos:broader 同 skos:narrower、skos:broaderTransitive 和 skos:narrowerTransitive 都互为逆属性, 而且后两者同 skos:related 不相交, 即两个概念间不能既存在等级关系, 又存在相关关系。

这里需要说明一下 skos:broader 和 skos:broaderTransitive 的区别和使用方式 (skos:narrower 和 skos:narrowerTransitive 类似)。除了上下位类关系外, skos:broaderTransitive 同 skos:broader 的不同之处在于, 前者被定义为 OWL 传递性属性, 而后者则不是。这意味着, 如果有以下声明:

A skos:broader B, B skos:broader C

则不能直接推理出:

A skos:broader C

但可以推理出:

A skos:broaderTransitive C

这样, skos:broader 和 skos:narrower 主要用于声明直接的上下位类关系, 而 skos:broaderTransitive 和 skos:narrowerTransitive 不用于直接声明上下位类关系, 而是用于记录推理得到的上下位类关系。

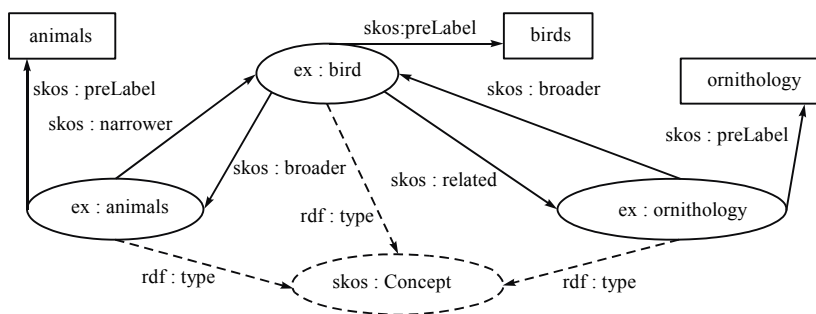


图 8.10 SKOS 语义关系属性的使用: 一个例子^①

相应地用 XML/RDF(S)语法表示如下。

```
<rdf:RDF
  xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos = "http://www.w3.org/2004/02/skos/core#" >
  <skos:Concept
    rdf:about = "http://www.example.com/concepts#birds" >
    <skos:prefLabel > birds </skos:prefLabel >
    <skos:broader rdf:resource =
      "http://www.example.com/concepts#animals"/ >
    <skos:related rdf:resource =
      "http://www.example.com/concepts#ornithology"/ >
```

① 图 3.16 中虚线部分不会出现在 RDF/XML 表示中, 下同。

```

</skos:Concept>
<skos:Concept
  rdf:about="http://www.example.com/concepts#animals">
  <skos:prefLabel>animals</skos:prefLabel>
  <skos:narrower rdf:resource=
    "http://www.example.com/concepts#birds"/>
</skos:Concept>
<skos:Concept rdf:about=
  "http://www.example.com/concepts#ornithology">
  <skos:prefLabel>ornithology</skos:prefLabel>
  <skos:related rdf:resource=
    "http://www.example.com/concepts#birds"/>
</skos:Concept>
</rdf:RDF>

```

⑥ 标签关系。

标签关系主要用于描述概念的标签之间的关系，具体包括以下建模元语：标签关系（skos:LabelRelation）、标签相关（skos:labelRelated）、参见标签关系（skos:seeLabelRelation），其中前者被定义为 OWL 类，后两者则分别被定义为 OWL 数据属性和 OWL 对象属性。

SKOS 中包含这些建模元语的目的是作为潜在的扩展点，允许 SKOS 用户自行定义更丰富的、精专的标签间关系，如两个标签间是不同语言翻译、缩写等关系。其中，skos:LabelRelation 用来声明一个资源为标签关系，skos:labelRelated 用于对标签关系进行注解，skos:seeLabelRelation 则用于引导到标签关系。

⑦ 概念集合。

当 SKOS 中的一些概念在某些方面有相似性时，用户可能需要将其集合起来使用，概念集合相关建模元语就出于这个目的，包括：概念集合（skos:Collection）、有序概念集合（skos:OrderedCollection）、集合成员（skos:member）、集合成员列表（skos:memberList）。其中，skos:Collection 和 skos:OrderedCollection 被定义为 OWL 类，后者是前者的子类；skos:member 和 skos:memberList 则被定义为 OWL 对象属性。skos:Collection 用来声明概念集合，skos:OrderedCollection 则用于声明有序的概念集合；与之对应，skos:member 和 skos:memberList 分别用于声明概念集合和有序概念集合中的元素。

⑧ 概念映射关系。

不同概念框架中的概念之间可能存在着内在的关系，为了互操作性，需要在不同概念框架中的概念间建立映射，SKOS 提供了相应的建模元语，包括：映射关系（skos:mappingRelation）、准确匹配（skos:exactMatch）、上位匹配（skos:broadMatch）、下位匹配（skos:narrowMatch）及相关匹配（skos:relatedMatch），它们都被定义为 OWL 对象属性，其中后四者是前者的子属性。顾名思义，skos:exactMatch 用于声明两个不同的概念足够相似，在信息检索等应用中彼此可以互替；skos:broadMatch 和 skos:narrowMatch 用于声明两个概念存在着等级映射关系，如一个概念被映射为另一个概念的上位类，它们之间互为逆属性；skos:relatedMatch 用于声明两个概念之间存在着相关关系，它同 skos:broadMatch/ skos:narrowMatch 是不相交的，即两个概念不能同时既有等级映射关系，又有相关映射关系。

2) SKOS 评价

SKOS 是一套建立在 RDF (S) 基础上的 NKOS 表示语言, 它简练而实用且有良好的扩展性, 非常适用于当前 NKOS 表示的主流需要。SKOS 的重要意义在于为当前受控词表的表示提供了一套推荐标准, 这将大大促进受控词表的编制与利用, 有着广阔的应用前景。SKOS 的主要问题在于它能表示的语义关系还比较有限, 远少于受控词表标准 Z39.19 中定义的数量——当然, 这个问题可以通过新增定义来轻易解决。另一个潜在问题是 SKOS 的推理机制还是比较薄弱的——虽然 SKOS 以表示受控词表等轻量级 NKOS 为主要目标, 但这种薄弱的推理能力可能在一定程度上影响 SKOS 的使用效果。无论怎样, 需要看到的是 SKOS 作为一种新近产生的语言标准, 尚处于不成熟阶段, 这就需要图书馆情报学界、计算机应用领域等相关领域的相关科研人员通过共同努力来完善它。

关于机器标识语言与编码, 图 8.11 概括了本节所涉及的 NKOS 表示语言之间的演化关系: 其中深灰色的为 W3C 推荐标准, 浅灰色的为目前已经很少使用的语言。虚线左侧为语言, 右侧为相应的逻辑机制。左侧半部分中, 虚线以上的语言具备了表示语义的能力, 虚线以下的语言适于表示本体。

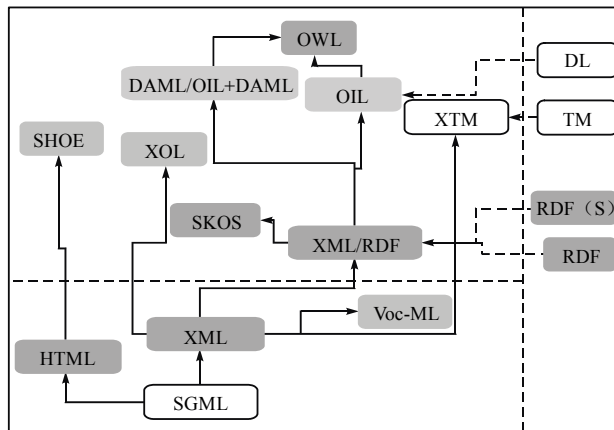


图 8.11 NKOS 表示语言概览

8.4.2 SKOS-XL 语言简介

为了能够对标签进行更严格的管理, SKOS 在其 eXtension for Labels (XL) 中增加了 skosxl:Label 类来弥补原定义的松散, 作为 skosxl:prefLabel, skosxl:altLabel 和 skosxl:hiddenLabel 属性的 Range。XL 还扩展定义了 skosxl:LabelRelation 属性来表示 skosxl:Label 实例之间的直接 (二元) 关系, 并建议以此作为扩展点提炼更专指的链接类型。但该属性的存在等于另建了一套语言标签之间的关系体系, 将概念间关系与语言标签间的关系割裂开来。这种松散的概念定义方式使 SKOS 拥有较好的表达能力, 可以用来表示各种规范程度的受控词表, 但用来表示高受控词表时, 无疑会丢失其本应拥有的推理能力, 不能实现严格的一致性检测^①。

SKOS-XL (表 8.12) 命名空间的 URI 是 <http://www.w3.org/2008/05/skos-xl#>, 前缀 skosxl: 是 SKOS-XL 的命名空间 URI 的缩写。

① SKOS Simple Knowledge Organization System Reference[EB/OL].(2009-08-18)[2013-06-10].<http://www.w3.org/TR/skos-reference/#xl>

表 8.12 SKOS-XL 的词汇表^①

URI	Defined by (http://www.w3.org/TR/skos-reference/#xl)
skosxl:Label	The skosxl:Label Class
skosxl:literalForm	The skosxl:Label Class
skosxl:prefLabel	Preferred, Alternate and Hidden skosxl:Labels
skosxl:altLabel	Preferred, Alternate and Hidden skosxl:Labels
skosxl:hiddenLabel	Preferred, Alternate and Hidden skosxl:Labels
skosxl:labelRelation	Links Between skosxl:Labels

8.4.3 SKOS 语言应用实例

1) AGROVOC 词表的 SKOS 表示

联合国粮食与农业组织 (Food & Agriculture Organization, 简称 FAO) 构建了一个知识组织系统的注册系统, 存储了 90 多个与农业和农业管理相关的知识组织系统, 其中包括非常有影响力的 AGROVOC 多语言农业词表。在该注册系统中, 能够按照词表的领域和类型浏览注册词表。

AGROVOC 是一部多语种结构的叙词表, 它涵盖了农业、林业、渔业、食物安全及其他相关学科领域 (例如: 可持续发展、营养学等)。它由词或词组 (术语) 组成, 含有不同语言, 具备各类词间关系 (例如: “广义”、“狭义”、“相关”等), 它主要用于标引或检索信息资源。它的主要作用是将信息标引标准化, 从而使得信息检索更加简单并且准确, 为用户提供最准确的信息资源。AGROVOC 是由联合国粮食及农业组织 (FAO) 和欧共体在 20 世纪 80 年代初开发的, FAO 大约每 3 个月进行一次维护更新, 用户可以在 AGROVOC 网址上看到更新变化。

此外, FAO 还针对 AGROVOC 多语言词表开发了一组基于 SOAP 协议的 Web 服务, 支持机器对 AGROVOC 词表内容的访问, 如表 8.13 所示^②。FAO 注册系统主页: [http://aims.fao.org/website/Knowledge-Organization-Systems-\(KOS\)/sub](http://aims.fao.org/website/Knowledge-Organization-Systems-(KOS)/sub)。

表 8.13 基于 AGROVOC 多语言农业词表的术语服务^③

<ul style="list-style-type: none"> • getAgrovocLanguages 获取 AGROVOC 词表中目前所有可用的语种;
<ul style="list-style-type: none"> • getTermcodeByTerm(String term) • getTermcodeByTermXML(String term, String format) 获取 AGROVOC 词表中某个指定术语的代码;
<ul style="list-style-type: none"> • getTermByLanguage(String termcode, String language) • getTermByLanguageXML(String termcode, String language, String format) 获取指定术语代码和指定语种的一个 AGROVOC 术语;
<ul style="list-style-type: none"> • getTermsListByLanguage2(String termcodes_list, String language, String separator) • getTermsListByLanguageXML(String termcodes_list, String language, String format) 获取指定术语代码和指定语种的一组 AGROVOC 术语;

① SKOS Simple Knowledge Organization System Reference[EB/OL].(2009-08-18)[2013-06-10].<http://www.w3.org/TR/skos-reference/#xl>

② Agrovoc Web Services Version 2.0 Documentation [EB/OL]. [2010-07-28]. <http://aims.fao.org/website/Documentation/sub>.

③ Agrovoc Web Services Version 2.0 Documentation [EB/OL]. [2010-07-28]. <http://aims.fao.org/website/Documentation/sub>.

续表

<ul style="list-style-type: none"> • getAllLabelsByTermcode2(String termcode, String separator) • getAllLabelsByTermcodeXML(String termcode, String format) <p>获取指定术语代码的一个 AGROVOC 术语的所有标签;</p>
<ul style="list-style-type: none"> • simpleSearchByMode2(String searchString, String searchmode, String separator) • simpleSearchByModeXML(String searchString, String searchmode, String format) <p>在指定搜索模式下, 搜索含有指定字符串的所有 AGROVOC 术语;</p>
<ul style="list-style-type: none"> • searchByTerm2(String searchString, String separator) • searchByTermXML(String searchString, String format) <p>获取含有指定字符串的所有 AGROVOC 术语;</p>
<ul style="list-style-type: none"> • searchCategoryByMode(String searchString, String language, String schemeid, String searchmode, String separator) • searchCategoryByModeXML(String searchString, String language, String schemeid, String searchmode, String format) <p>在指定搜索模式下, 搜索某个指定分类体系中含有指定字符串的所有类目;</p>
<ul style="list-style-type: none"> • simpleSearchByModeandLang(String searchString, String searchmode, String separator, String language) • simpleSearchByModeLangXML(String searchString, String searchlanguage, String searchmode, String format, String outputlanguage) <p>在指定搜索模式下, 搜索含有某个语种的指定字符串的指定语种的所有 AGROVOC 术语;</p>
<ul style="list-style-type: none"> • getRelationByTermcodeXML (String termcode, String relation, String format) <p>获取和指定术语代码的术语具有指定关系的一个 AGROVOC 术语;</p>
<ul style="list-style-type: none"> • getStatus (String termcode, String language) <p>获取具有指定术语代码和指定语种的术语的状态;</p>
<ul style="list-style-type: none"> • searchByModeLangScopeXML(String searchString, String language, String searchmode, String format, String outputlanguage, String scopeid) <p>在指定搜索模式下, 搜索指定语种和指定范围内含有某个语种的指定字符串的 AGROVOC 术语;</p>
<ul style="list-style-type: none"> • getConceptInfoByTermcode (String termcode) • getConceptInfoByTermcodeXML(String termcode, String format) <p>获取具有指定术语代码的一个 AGROVOC 术语的重要概念信息, 即所有语种的标签、广义概念、狭义概念、相关概念、非首选标签等;</p>
<ul style="list-style-type: none"> • getDefinitions(int termcode, String language) • getDefinitionsXML(int termcode, String language, String format) <p>获取具有指定术语代码和指定语种的术语的定义、历史注释和范围注释;</p>
<ul style="list-style-type: none"> • getTermExpansion (String term, String language) <p>获取一个指定 AGROVOC 术语的某个指定语种的所有同义词;</p>
<ul style="list-style-type: none"> • complexSearchByTermXML (String searchStrings, String separator, String format) <p>搜索以指定的一组字符串起始的所有 AGROVOC 术语。</p>

2) LCSH 词表的 SKOS 表示

LCSH 是 Library of Congress Subject Headings 的简称, 是美国国会图书馆编制的一部大型综合性标题表。LCSH 是目前世界上规模最大、应用最广泛的标题表, 在检索语言的发展史上和当今图书馆主题编目工作中, 占有重要地位。美国国会图书馆的主题标目是以机器可处理的 MARC 形式存在的, 近来已转为 MARCXML 编码形式。美国国会图书馆的 MARC 服务拥有 26.5 万条 LCSH 规范记录, 传统上以 MARC21 格式进行发布。以下将以 LCSH/MARC 指代“国会图书馆主题标目”, 特指以 MARCXML 形式存在的机读数据, 并以 LCSH/SKOS 指代以 SKOS 形式编码的 LCSH 数据。

(1) 以 SKOS 表示 LCSH。

① 基本模型。

Harper^①最早对 LCSH/MARC 向 SKOS 的转换进行了大量的探索,并提供了具体的 XSLT 转换代码,能够将 MARCXML 规范数据映射为 SKOS。SKOS 和 LCSH/MARC 都有自己的概念模型, LCSH/MARC 将不同的标目(规范/非规范)聚集为记录,成为表达语义关系和附注的抽象的概念实体。类似地, SKOS 词表也主要是由 skos: Concept 实例构成,成为以 URI 标识的“思想单元”,并且 SKOS 的概念也都有词汇标签和说明文本,以及指向其他概念的丰富的语义链接。

② 概念。

每个 LC 的 MARC 规范记录都在 001 字段著录有国会图书馆控制码(LCCN),这让其成为标识 SKOS 概念的最好候选。LCCN 具有永久性和唯一性的特点, SKOS 需要 URI 来标识 skos: Concept 实例。正如 2004 年 Frank Manola 等人的文章中所指出的,语义网技术所采用的 RDF 以及“链接数据”所推荐的 HTTPURL 方式都需要用 URI 来标识资源,这样才能使资源更容易地被获取^②。当然, LCCN 毕竟不是 URL,还需要通过模板: `http://lcsch.info/{lccn}#concept` 进行规范化,并加入到 URI 中。

以 LCCN 作为 URI 的做法与 Harper^③提出的方法稍有不同, Harper 是以规范标题文本构建 URL,如这种形式: `http://example.org/World+Wide+Web`。论文的作者一般都喜欢用 LCCN 作为 URI 标识,因为标题总是在发生变化,而 LCCN 记录号一般都能保持相对稳定。万维网以及当前热门的语义万维网实践都建议 URI 应该相对稳定,尽可能不要随时间而变化^④。由于概念的标识采用了永久性的 URL,也使得 LCSH/SKOS 概念的元数据描述具有了持久性。

③ 词汇标签。

MARC21 规范数据格式规定了规范标题(1xx)和非规范标题(4xx)的区别,相应地, SKOS 词汇表也提供了两个属性元素: skos: prefLabel 和 skos: altLabel,这使得一个概念可以对应两种自然语言表达:首选项和交替项。也就是说,这样就能够使 LCSH 的规范标目和非规范标目分别直接映射到 skos: prefLabel 和 skos: altLabel 属性上。

但是即使这样,仍然有大量信息丢失。MARC 用以著录规范标目的字段还包括时序(Chronological, MARC21 的 148 字段)、主题(Topical, 150 字段)、地理(Geographic, 151 字段)、体裁/格式(Genre/form, 155 字段)等,对于 LCSH/SKOS 来说,能够获取这些概念的描述也十分重要。

LCSH/MARC 还有大量的规范标目来自于其他标目的合并,也就是常说的“先组”(Pre-coordination)方式。举个例子,一个主题标目“Drama”可以与一个时序标目“17th century”进行组合,形成规范标目 Drama-17th century,通过主副标目的“分面”应用,信息被清晰的表示。而在 LCSH/SKOS 表示中,标目之间仅是一种平面的文字关系,即“Drama-17thcentury”。这是 SKOS 需要扩展的一个地方。

① Harper, Corey. Authority Control for the Semantic Web. Encoding Library of Congress Subject Headings. International Conference on Dublin Core and Metadata Applications[EB/OL], Manzanillo, Mexico. [2008-06-20].<http://hdl.handle.net/1794/3268>.

② Sauermann, Leo, Richard Cyganiak. Cool UR Is for the Semantic Web[EB/OL].[2008-06-20]. <http://www.w3.org/TR/cooluris/>.

③ Harper, Corey. (2006). Authority Control for the Semantic Web. Encoding Library of Congress Subject Headings. International Conference on Dublin Core and Metadata Applications[EB/OL], Manzanillo, Mexico. [2008-06-20]<http://hdl.handle.net/1794/3268>.

④ Berners-Lee, Tim. CoolUR Is don't Change. [EB/OL].[2008-06-20]. <http://www.w3.org/Provider/Style/URI>

SKOS 是为多语言环境而设计的, SKOS 鼓励用户使用语言标签^①来标识语种, 例如:

```
ex: animals rdf: type skos: Concept;
skos: prefLabel " animals" @en;
skos: prefLabel " animaux" @fr.
```

④ 语义关系。

LCSH/MARC 使用 5xx 字段连接相关的规范标目。SKOS 则通过 skos: related, skos: broader, skos: narrower 等元素提供丰富的概念资源之间的语义联系。

LCSH/MARC 里的语义关系是很容易转换为 LCSH/SKOS 的。LCSH/MARC 采用标目建立参考链接, 而 LCSH/SKOS 使用概念资源的 URI 建立相互联系。转换程序要为被转换的特定标题寻找 URI, 以建立链接关系。除此之外, LCSH/MARC 没有标注下位类关系, 只是通过上位类关系来缺省表达, 因此建立 skos: broader 链接时, 需要同时创建明确的 skos: narrower 属性联系。一旦完成了用 URI 标识概念资源, 就会形成类似图 8.12 一样的结构, 图 8.12 描述了与概念“World Wide Web”相关的概念。

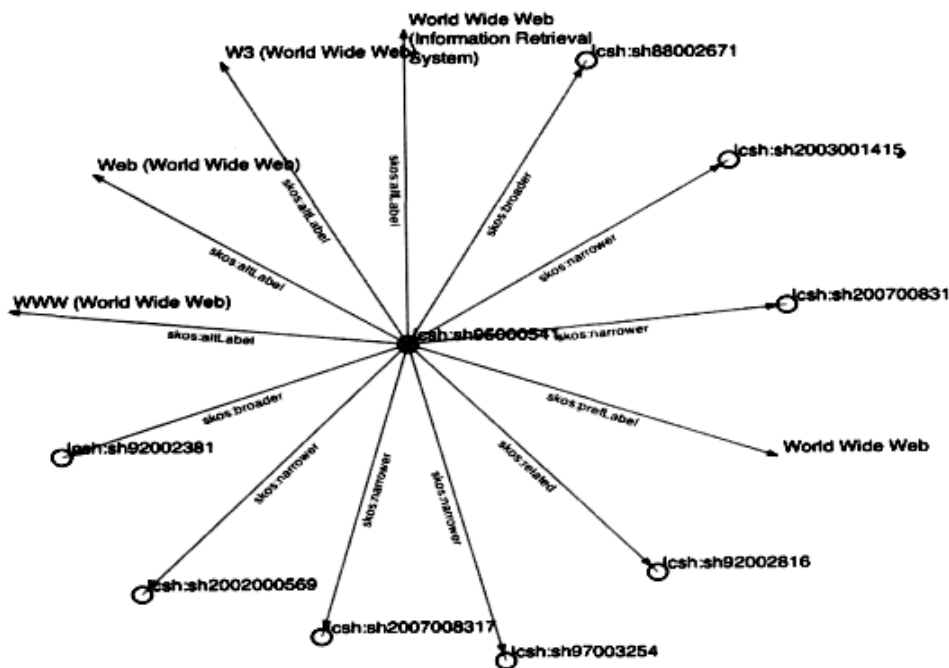


图 8.12 SKOS 概念图示

⑤ 文件属性。

LCSH/MARC 中有许多有关标题的说明性字段, 例如: 说明 (General note, 667) 字段、数据来源 (Sourcedata, 670) 字段、举例 (Examples, 681) 字段等。SKOS 词表也包括一些说明性属性可以用于 LCSH/SKOS 编码, 如 skos:note, skos:editorialNote, skos:definition, skos:scopeNote, skos:changeNote, skos:historyNote 等。这些属性只需稍加琢磨, 就能应用于将 LCSH/MARC 转换成 LCSH/SKOS。

① Isaac, Antoine, Ed Summers. SKOS Simple Knowledge Organization System Primer [EB/OL]. [2008-06-20]. <http://www.w3.org/TR/skos-primer/>.

⑥ 使用非 SKOS 的说明性属性。

LCSH /MARC 包含一些其他 SKOS 词汇本身所缺乏的特性, 比如国会图书馆分类号范围、记录创建日期、记录最后修改日期等。然而由于 RDF 的灵活性, 允许使用其他词汇, 如都柏林核心元数据的术语: `dcterms:lcc`, `dcterms:created`, `dcterms:modified` 等, 混搭到 SKOS 描述中, 以获取这些属性信息。混搭使用其他 RDF 词汇当然也可以, 这种不需要事先进行编码模式定义的灵活性是采用 RDF 的一个强大而吸引人的特性。

⑦ LCSH /SKOS 映射。

文中涉及的转换可以总结成如表 8.14 所示的字段元素映射表。

⑧ LCSH /SKOS 插图。

一旦一个 LCSH /MARC 记录被转换成 LCSH /SKOS, 即创建了一组 RDF 表示, 参见图 8.13。这是一个关于“World Wide Web”概念的例子。LCSH /MARC 记录之间的文本链接已转换为资源之间明确的 URI 链接(参见图 8.13)。

表 8.14 字段元素映射表

MARC Field	Feature /Function	RDF Property	Value of the Property/Comments
010	Control Number	<code>rdf:about</code>	the URI for the skos:Concept instance
150	Topical Term	<code>skos:prefLabel</code>	subfields: a, b, v, x, y, z
151	Geographic Term	<code>skos:prefLabel</code>	subfields: a, b, v, x, y, z
450	See From Tracing(Topical Term)	<code>skos:altLabel</code>	subfields: a, b, v, x, y, z
451	See From Tracing(Geographic Name)	<code>skos:altLabel</code>	subfields: a, b, v, x, y, z
550	See Also From Tracing(Topical Term)	<code>skos:broader</code>	only use this property when subfield w is 'g'; use value to lookup ConceptURI
550	See Also From Tracing(Topical Term)	<code>skos:related</code>	only use this property when subfield w is not present with 'g' or 'h' in position 0; use value to lookup ConceptURI
551	See Also From Tracing(Geographic Name)	<code>skos:broader</code>	only use this property when subfield w is 'g'; use value to lookup ConceptURI
551	See Also From Tracing(Geographic Name)	<code>skos:related</code>	only use this property when subfield w is not present with 'g' or 'h' in position 0; use value to lookup ConceptURI
667	Non Public General Note	<code>skos:note</code>	subfield: a
670	Source Data Found	<code>dcterms:source</code>	subfields: a, b, u
675	Source Data not Found	<code>skos:editorialNote</code>	subfield: a
678	Biographic or Historical Data	<code>skos:definition</code>	subfields: a, b, u
680	Public General Note	<code>skos:scopeNote</code>	subfields: a, i
681	Subject Example Tracing Note	<code>skos:example</code>	subfields: a, i
682	Deleted Heading Information	<code>skos:changeNote</code>	subfield: a, i
688	Application History Note	<code>skos:historyNote</code>	subfield: a
008	Fixed Length Data Elements	<code>dcterms:created</code>	positions 0 - 5
005	Date and Time of Last Transaction	<code>dcterms:modified</code>	
053	LC Classification Number	<code>dcterms:lcc</code>	subfield: a

3) 汉语主题词表的 SKOS 表示

《汉语主题词表》中的术语采用表 8.15 中所示的 27 个属性进行描述, 其中某些属性是为创建词表而设立的(如增词时间、词频、词类型、浏览次数、编辑次数等), 无需向用户展示, 因此予以忽略, 其余属性则需转换为相应的 SKOS 属性。如果 SKOS 标准语言中没有与词汇属性相对应的属性, 我们则对 SKOS 标准语言进行定制化扩展, 增加新属性, 扩展部分称为 SKOEX 语言。无论是 SKOS 标准语言中固有的属性还是扩展语言中新增的属性都统称为 SKOS 属性。表 8.15 中列出了《汉语主题词表》中的词汇属性与 SKOS 属性之间的映射关系, 并在下文中给出了详细说明。

(1) 00-中图分类号, 01-范畴。

在叙词表中, 同一个分类号或范畴号一般对应于多个叙词概念, 并不是叙词所唯一具有的某种标记符号。因此, 叙词表中叙词概念所对应的分类号或范畴号不采用 `skos:notation` 属性来描述, 而是采用映射属性 `skos:mappingRelation`, 此时, 对应的分类法或者范畴索引分别看作是独立的概念体系。因为叙词概念是归属于分类体系或范畴体系的类目概念之下

的，因此确切地说是采用映射属性的子属性 `skos:broadMatch` 来描述。

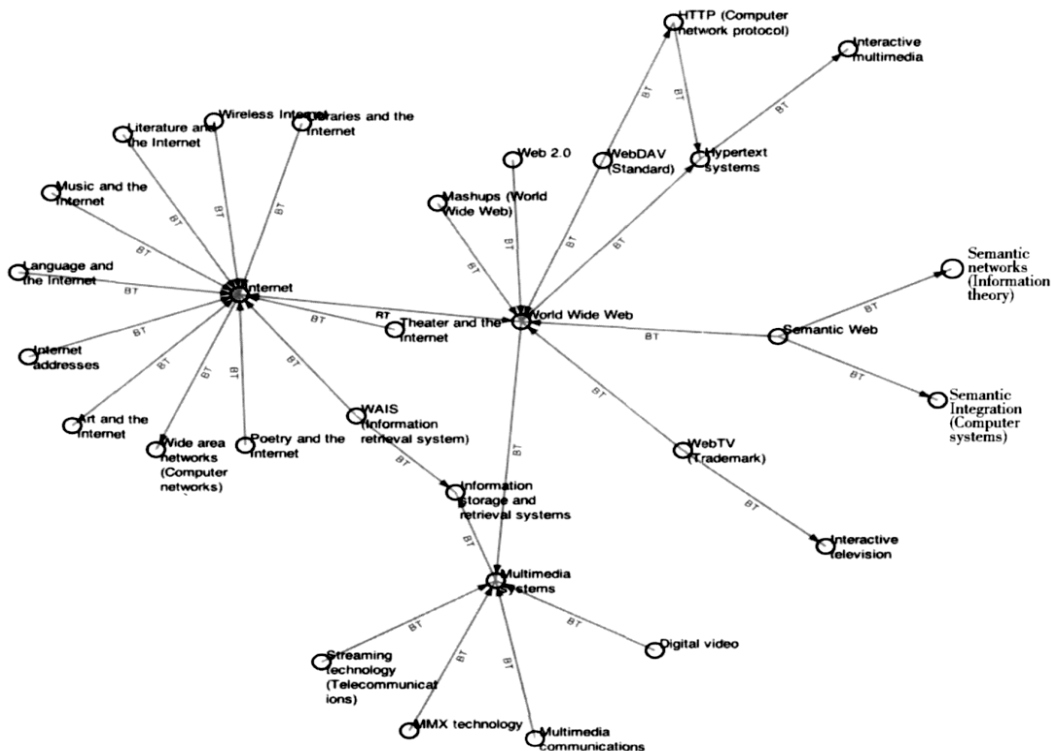


图 8.13 概念间的语义关系

(2) 02-拼音, 03-英文, 19-中文, 21-缩略语。

这四个属性都表示叙词概念的标签。中文、英文和拼音标签分别作为相应语种的首选标签，采用属性 `skos:prefLabel` 表示，语种则分别采用属性 `xml:lang="zh/en/zh-Latn"` 表示。有极少数叙词具有多个英文标签，此时只能将一个英文标签作为英文语种的首选标签，其他的均作为可选标签。缩略语是可选标签的一种，采用属性 `skos:altLabel` 的新增子属性 `skosex:abbreviation` 来表示。

(3) 05-代项。

该属性只用于叙词，指与叙词同义的非叙词，该非叙词则作为对应叙词的一个可选标签，采用 `skos:altLabel` 属性描述。

(4) 04-用项。

该属性只用于非叙词，指与非叙词所对应的叙词，两者是同义关系。该属性与“05-代项”是互逆关系。

表 8.15 《汉语主题词表》中的词汇属性与 SKOS 属性的映射关系

属性代码	属性	对应的 SKOS 属性	说明
00	中图分类号	<code>skos:broadMatch</code>	在《中图法》中所属的类号
01	范畴	<code>skos:broadMatch</code>	在来源词表中所属的范畴号 (CT_SCI)
02	汉语拼音	<code>skos:prefLabel</code> <code>xml:lang="zh-Latn"</code>	叙词的拼音表示

续表

属性代码	属性	对应的 SKOS 属性	说明
03	英文	skos:prefLabel xml:lang="en"	叙词的英译名称
04	用项		与非叙词同义的叙词
05	代项	skos:altLabel xml:lang="zh"	与叙词同义的非叙词
06	属项	skos:broader	上位概念
07	分项	skos:narrower	下为概念
08	参项	skos:related	相关概念
09	族项	skosex:topBroader	族首词, 即词族中最上位的概念
10	组面	skosex:facet	在分面分类法中, 用某一单一系列的分类标准对一个主题领域进行划分而产生的一组类目
11	注释	skos:note	注释属性
12	用和	skosex:coordinationOf	组配生成复合概念的成份概念
13	组代	skosex:coordinatedTo	单一叙词组配而成的复合概念。
14	领词	skosex:leadBroader	分词族的族首词
15	代码	skos:notation	叙词对应的某种标记符号
16	和项	skosex:coordinationOf	组配生成复合概念的一个成份概念
17	历史注释	skos:historyNote	历史注释
18	同项	skos:exactMatch	两个叙词之间的等同关系
19	中文	skos:prefLabel xml:lang="zh"	叙词的中文标签
20	增词时间		忽略
21	缩略语	skosex:abbreviation	缩略语也是叙词可选标签的一种
22	词频		忽略
23	浏览次数		忽略
24	编辑次数		忽略
25	用户评价	skosex:evaluationNote	用户评价注释
26	词类型		忽略

注: (1) xmlns:skos="http://www.w3.org/2004/02/skos/core#" 表示 W3C 定义的 SKOS 标准语言;

(2) xmlns:skosex="http://www.example.com/2011/06/skos/extension#" 表示对 SKOS 扩展语言;

(3) 属性 xml:lang="xx" 中的语言代码由《IETF BCP 47》标准定义

(5) 06-属项, 07-分项, 08-参相。

这三个属性分别用来描述两个叙词概念之间的上位关系、下位关系和相关关系, 分别对应 skos:broader, skos:narrower 和 skos:related 属性。

(6) 09-族项。

新增属性 skosex:topBroader 作为 skos:broader 属性的子属性, 用以描述叙词与其所在词族等级中族首词间的关系。之所以不采用 skos:broader 属性来描述, 因为该属性描述的是所有上位关系, 并不专指词族中的顶级概念。

(7) 14-领词。

当一个词族太大时, 将比族首词小一级或二级且下位概念较多的叙词分出来构成分词族, 分词族的族首词称领词。领词是词族中位于某个中间层级的概念, 是族首词的一个下位

概念, 是所描述的叙词的一个上位概念。新增属性 `skosex:leadBroader` 作为 `skos:broader` 属性的子属性, 用以描述叙词与领词之间的关系。

(8) 10-组面。

在分面分类法中, 对类目上位类进行分面分析, 形成几组不同性质的下位类, 每一组称作一个组面。因此组面是用某一单一系列的分类标准对一个主题领域进行划分而产生的一组类目, 即表示某一类事物某一方面属性的一组简单概念。一个组面可看作是一组下位类, 组面标签本身用新增属性 `skosex:facet` 表示, 作为属性 `skos:narrower` 的一个子属性。

(9) 12-用和, 13-组代, 16-和项。

将两个或两个以上已存在的叙词概念组配起来, 代替一个词表中尚未存在的复杂概念, 组配叙词与这个复杂概念之间的关系称为组代参照, 采用新增属性 `skosex:coordinatedTo` 表示。“用和”表示组配生成的复合概念的成份概念, “和相”则表示另一个成份概念, 均采用新增属性 `skosex:coordinationOf` 来表示。

(10) 15-代码。

指叙词所唯一具有的某种标记符号, 因此采用 `skos:notation` 属性描述, 两者之间是一一对应关系。

(11) 18-同项。

采用 `skos:exactMatch` 属性关联两个具有相同含义的叙词概念。之所以不采用 `owl:sameAs` 属性, 是因为采用该属性链接的两个资源被认为是同一个资源, 因此它们的 RDF 三元组是可以合并的, 这会导致同一个概念对于同一个语种有两个首选标签。

(12) 11-注释, 17-历史注释, 25-用户评价。

用户评价也是对叙词的一种注释, 因此新增属性 `skos:note` 的子属性 `skosex:evaluationNote` 来表示。注释和历史注释分别用 `skos:note` 和 `skos:historyNote` 表示。

(13) 20-增词时间, 22-词频, 23-浏览次数, 24-编辑次数, 26-词类型。

这些属性是在词表构建中所使用的属性, 主要是为词表构建者提供某种选词信息, 不作为面向用户的属性使用, 因此在进行语义化转换时予以忽略。

图 8.14 为采用 SKOS/SKOS-XL 及其扩展语言 SKOSEX 描述的《汉语主题词表》中“情报检索”一词及其相关概念的 RDF 图。在实际应用中, 采用 RDF/XML 序列化格式进行表示。

8.4.4 术语注册与术语服务

术语表、分类表、叙词表、本体等各种词表 (即知识组织系统)^① 在信息资源描述、组织、管理、发现等方面的强大功能已经得到图书情报界和相关领域的广泛认可。为促进对这些知识组织工具的有效利用, 需要对它们进行组织和管理。早期的做法是在机构内部创建和维护各种印刷版本的词表列表以供用户使用, 如欧盟发布的 *Thesaurus Guide*^②。自 1996 年起国外陆续出现了一些以电子格式发布的在线词表列表, 如英属哥伦比亚大学图书情报学院的词表索引^③和 *HILT Resource List*^④, 遗憾的是这些列表中的大多数并没有得到持久的扩展和维护。20 世纪 90 年代末网络知识组织系统 (Networked Knowledge Organization Systems/

① 本文中的词表均指广义的词表, 与知识组织系统等意, 可互换使用。

② EUROBroker S. *Thesaurus guide: Analytical directory of selected vocabularies for information retrieval*, 1992 (2nd version) [R]. Luxembourg: European Communities, 1993.

③ HILT vocabulary resources [EB/OL]. [2011-01-28]. <http://hilt.cdli.strath.ac.uk/Sources/vocabulary.html>.

④ Stephenson M. *Indexing resources on the WWW: database indexing, controlled vocabularies & thesauri* [EB/OL] [2011-01-28]. <http://www.slais.ubc.ca/resources/indexing/database1.htm>.

Services, 简称 NKOS2) 社区开始了研制术语注册的尝试, 知识组织资源的存储、组织、管理和利用开始朝着有序化、规范化和网络化的方向发展。

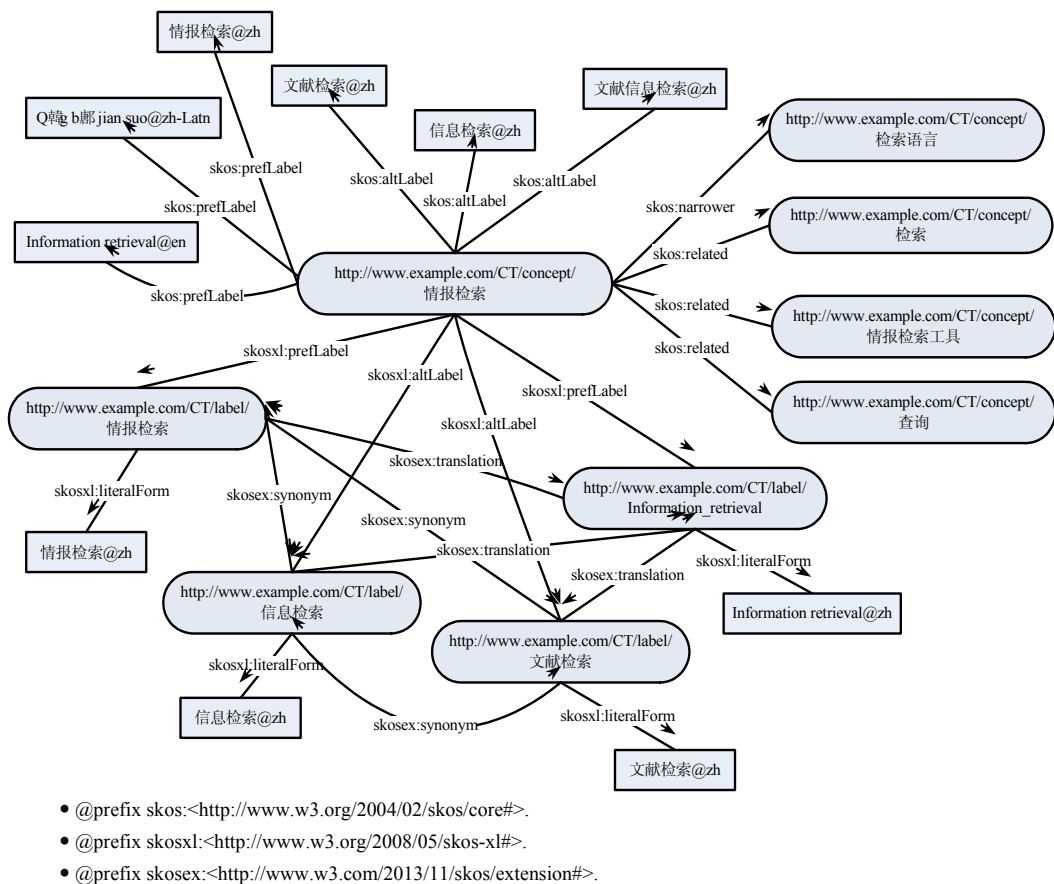


图 8.14 采用 SKOS/SKOS-XL 及其扩展语言 SKOSEX 描述的“情报检索”一词及其相关概念

术语注册是指对各种词表提供权威的、集中控制的存储, 以促进词表的发现、重用、管理、标准化和互操作。一个术语注册系统能够列出、描述、识别并且指明在信息系统和信息服务中可用的词表集合, 并且提供图形化界面和术语服务以供用户访问和使用词表内容(指词表成员术语、概念及其相互关系)^①。所谓术语服务是指对词表元数据和词表内容进行浏览、查询、应用的各种 Web 服务的统称^②。术语服务通过 Web 应用程序接口(API)支持机器对词表及其内容的访问和调用, 是在网络环境下对词表进行应用的重要途径。术语注册和术语服务两者相辅相成, 前者是后者的前提和保证, 后者是前者的目的和应用。

国外代表性的术语注册系统有 Taxonomy Warehouse^③、Lexaurus Bank^①、FAO VEST

① Golub K, Tudhope D. Terminology registry scoping study (TRSS): Final report [R/OL]. UK: Joint Information Systems Committee (JISC), 2009.

[2011-01-28]. <http://www.jisc.ac.uk/media/documents/programmes/sharedservices/trss-report-final.pdf>.

② Tudhope D, Koch T, Heery R. Terminology services and technology: JISC state of the art review [R/OL]. UK: Joint Information Systems Committee (JISC), 2006. [2011-01-28]. http://www.jisc.ac.uk/Terminology_Services_and_Technology_Review_Sep_06.

③ Taxonomy warehouse [OL]. [2011-01-28]. <http://www.taxonomywarehouse.com/>.

Registry、Open Metadata Registry、OCLC 术语服务等。Taxonomy Warehouse 是由 Dow Jones Factiva1 在 2001 年构建的 taxonomy 注册系统,共收集了由 288 个出版商提供的 670 个 Taxonomies,是最早建立的术语注册之一,但功能有限,只提供词表的分类浏览和名称检索。Lexaurus Bank 是英国 Vocabulary Management Group 公司开发的一个词表管理系统,支持 SKOS、Zthes、IMS VDEX2 等格式的词表的输入、输出以及分布式环境下词表的在线创建、编辑和相互映射,能够自动跟踪词表的更新修改并对词表进行完全的版本控制,此外还提供 REST 模式的 Web 服务以支持机器对机器的词表访问。FAO VEST Registry 是联合国粮农组织建立的一个综合性注册系统,词表大类中存储了 90 多个与农业和农业管理相关的词表,提供基于词表类型和领域的词表浏览,此外还针对 AGROVOC 多语言农业词表提供了一组基于 SOAP 协议的术语服务,实现对该词表中术语及其关系的检索。Open Metadata Registry 是在美国自然科学数字图书馆研究项目中构建的一个大型词表和元数据注册系统。是目前最强大的术语注册系统,不仅拥有基本的词表元数据和词表内容检索功能,还支持词表的在线编辑和更新、词表的版本控制、词表更新的自动通知等复杂功能,但遗憾的是该系统目前主要是通过可视化图形界面供人类用户使用,还没有提供支持机器访问的术语服务,开发者拟在后序工作中实现^②。OCLC 术语服务是 OCLC 开发的一个实验性术语服务系统,目前存储了包括 LCSH 在内的六个词表,支持 HTML、MARC XML、Zthes 和 SKOS 四种词表表示格式,采用 SRU 检索协议和 CQL 查询语言实现了一组术语服务^③。目前 OCLC 的术语服务已有了一些实验性的应用,譬如美国印第安纳大学的 OPAC 系统采用 OCLC 术语服务提供了一个查询扩展功能。

除上述专门的术语注册和服务系统外,在一些相关研究项目中也涉及到了术语服务的研究和开发,如 HILT、STAR 和 ADL 叙词表协议。HILT (High-level Thesaurus) 是英国 JISC (联合信息系统委员会)和 RSLP(研究支持图书馆计划)共同资助的一个研究项目,采用 SOAP 协议和 SRU/SRW 协议实现了七个用于术语检索的术语服务,检索结果以 SKOS 格式表示。STAR (Semantic Technologies for Archaeological Resources) 是英国 AHRC (艺术与人文研究委员会)的一个研究项目,采用 SKOS 为词表的表示格式,以 SKOS API 为词表内容的查询接口,开发了七个术语服务,提供术语查找、相关概念获取、概念扩展等功能。美国亚利山大数字图书馆项目中构建的 ADL 叙词表协议采用自定义的 XML 格式表示词表,提供了五个术语服务实现词表的查询和浏览,但是不支持词表的创建、维护、共享和相互映射等复杂操作^④。

目前术语注册中采用的词表表示格式主要是 XML 编码格式,但也有极个别系统支持 HTML 等非 XML 编码格式,如 OCLC 术语服务。XML 格式又可进一步细分为自定义 XML 格式和标准 XML 格式。自定义格式因不具有通用性,只在少数系统中出现,如 ADL 叙词表协议,大部分系统采用的是标准 XML 格式,主要有 MARC XML、Zthes 和 SKOS 三种。

MARC XML 是由美国国会图书馆制定的 MARC 21 格式的一种 XML 表示方式,是最早期的词表电子化表示格式。Zthes 被称作 Z39.50 词表描述模型,是一个基于 XML 格式的词

① Lexurus bank [OL]. [2011-01-28]. <http://www.vocman.com/lexaurusbank>.

② Hillmann D, Sutton S, Phipps J, et al. A metadata registry from vocabularies up: The NSDL registry project [C/OL]. // Baker, T. & Solorio, J. Proceedings of 2006 International Conference on Dublin Core and Metadata Applications: metadata for knowledge and learning. Colima, Mexico: Dublin Core Metadata Initiative. 2006: 65-75. [2011-01-28]. <http://arxiv.org/ftp/cs/papers/0605/0605111.pdf>.

③ Vizine-Goetz D. Terminology services [OL]. [2011-01-08]. <http://www.oclc.org/research/presentations/vizine-goetz/cendi-nkos-isko.ppt>.

④ Janée G, Ikeda S, Hill L. The ADL thesaurus protocol (version 1.0) [OL]. [2011-01-28]. <http://www.alexandria.ucsb.edu/thesaurus/specification.html>.

表描述和传输规范。这两种词表表示格式都是在较早时期制定的,目前已经不能适应网络环境下对词表应用的要求^①。SKOS 则是一种适用于网络环境下词表应用的新的表示格式。采用 SKOS 语言表示的词表能为机器可读可理解,适于词表在网络环境中下的应用,而且也更容易实现不同词表间的互操作,因此推荐在术语注册与服务系统中采用 SKOS 表示词表。

目前术语注册和服务系统中采用的协议和技术主要有以下几种:

(1) Web Service: Web 服务是基于 XML 和 HTTP 的一种服务,其通信协议主要基于 SOAP,采用 WSDL 对 Web 服务进行描述,通过 UDDI 来发现和获得服务的元数据。术语服务通常采用 Web 服务的方式进行发布和调用。

(2) SOAP (Simple Object Access Protocol): 是一个严格定义的信息交换协议,用于在 Web 服务中把远程调用和返回封装成机器可读的格式化数据。SOAP 数据使用 XML 数据格式,定义了一整套复杂的标签,用以描述调用的远程过程、参数、返回值和出错信息等。SOAP 可以和现存的许多互联网协议和格式结合使用,包括超文本传输协议 (HTTP),简单邮件传输协议 (SMTP),多用途网际邮件扩充协议 (MIME)。SOAP 协议是实现 Web 服务的一种比较成熟的方式。

(3) REST (Representational State Transfer): REST 是一种只使用 HTTP 和 XML 进行基于 Web 通信的技术。目前在三种主流的 Web 服务实现方案中,因为 REST 模式的 Web 服务与复杂的 SOAP 和 XML-RPC 对比来讲明显的更加简洁,越来越多的 Web 服务开始采用 REST 风格设计和实现。

(4) Z39.50: Z39.50 是一种在客户/服务器环境下计算机与计算机之间进行数据库检索的通信协议。它的出版及使用,解决了不同系统间的数据交流的问题,克服了信息检索网络化的障碍。

(5) Zing (Z39.50-International: Next Generation): Zing 是下一代的 Z39.50 协议,一方面可以看成是 Z39.50 各种功能在新的网络协议和应用模式下的拆解,另一方面也是一种简化。Zing 目前的版本是 1.0,包括 SRW/SRU、CQL、ZOOM、ez3959 和 ZeeRex 五个部分。

(6) SRW/SRU (Search/Retrieve for the Web 和 Search/Retrieve URL Service): SRU/SRW 是针对 Web 的信息检索协议,利用 Web 服务的架构实现了 Z39.50 的一些基本服务,是 ZING 的核心功能。SRW 使用 HTTP 与 SOAP 的无状态通信,采用 XML 作为信息传输编码,也可以单纯使用 URL 传递查询请求,用 WSDL 来定义 Z39.50 传输的格式信息,检索结果也以 XML 格式输出。而 SRU 只能通过 URL 参数方式提交检索请求,不支持完整的 SOAP 消息包(只支持 SOAP 消息包中的内容序列)。在 SRU 和 SRW 中采用 CQL 查询语言描述查询。

(7) SKOS API: 是对 SKOS 数据进行读写操作的程序接口,对采用 SKOS 表示的词表内容进行查询,可采用 SKOS API 进行。

(8) RDF API: 是对 RDF 数据进行读写操作的程序接口;对采用 RDFS 表示的本体内容进行查询,可采用 RDF API 进行。

(9) OWL API: 是对 OWL 数据进行读写操作的程序接口,对采用 OWL 表示的本体内容进行查询,可采用 OWL API 进行。

(10) SPARQL: 是针对 RDF 数据的一种标准查询语言,SKOS 版词表的内容数据采用 RDF/XML 格式表示,除了可采用 SKOS API 进行查询外,也可采用 SPARQL 语言进行查询。

因为 SKOS 格式逐步成为词表的主流表示格式,因此术语注册和服务系统多采用 SKOS API 对 SKOS 数据进行检索,采用基于 SOAP 的 Web 服务实现术语服务,但是近年来基于 REST 的 Web 服务也有所增多。

① 王军,张丽.网络知识组织系统的研究现状和发展趋势[J].中国图书馆学报,2008(1):65~69.

8.5 语义网信息组织方法

8.5.1 关联数据简介

关联数据是由万维网的创始人 Tim Berners-Lee 于 1996 年在他的“Design Issues for the World Wide Web”笔记中首次提出的一个概念，是指通过可解引用的 URIs（Dereferenceable URIs）地址在 Web 上展示、共享、连接数据的一种方式^①。关联数据的两个基本宗旨是：①采用 RDF 数据模型在 Web 上发布结构化数据；②采用 RDF 链接连接来自不同数据源的数据。关联数据必须遵循以下四个基本原则^②：

- 使用 URI 标识符命名任何事物；
- URI 标识符需是 HTTP URI 地址，使任何人都可以访问这些名称标识；
- 当有人访问某个标识名称时，采用 RDF、SPARQL 等标准提供有用的信息；
- 包含指向其他 URI 地址的链接，使人们可以发现更多相关事物。

在传统的文档 Web 中，对结构化数据的访问主要是通过 Web 应用程序接口（即 APIs）来实现。这种方式的缺点是不同的数据源一般都采用不同的访问界面，而且无法在来自不同数据源的数据之间设置超链接，因此各个数据源是孤立地存在着。虽然通过 Mashup^③ 可以合并若干个数据源，然后统一提供服务，但是这种方式也只能在有限的数据源中进行，无法扩展到整个 Web。关联数据提供了在 Web 上发布和访问结构化数据的一种新方式，是一种推荐的语义网最佳实践。通过这种方式，如同访问 Web 文档一样，能够直接通过 HTTP 协议访问结构化数据并且可以沿着数据间的链接在不同数据源间穿行，将所有数据构成一张数据之网。此外，相比于 Web 文档之间的超链接，数据之间的 RDF 链接更能够指明数据间的语义关系，有益于人机理解语境信息。与传统的基于应用程序接口的数据访问方式相比，关联数据提供了一种统一的、标准的数据访问机制，避免了访问界面和结果格式的纷繁复杂。通过采用关联数据，数据源更容易被搜索引擎抓取，能够采用通用的数据浏览器（即 RDF 浏览器）访问不同的数据源，能够在来自不同数据源的数据间建立链接^④。

8.5.2 关联数据中资源的命名及访问机制

在关联数据中，所有实体对象或抽象概念（如文献资源、个人、组织机构、地点、事件、术语等）都必须采用唯一的 HTTP URI 标识符进行命名，但是它们的 URI 地址不能够被 HTTP 协议直接解引用。它们在 Web 架构中被称为非信息资源，以区别于传统文档 Web 中 URI 地址能够被 HTTP 协议直接解引用的信息资源（如网页、图片或其他数字媒体格式等）。对于非信息资源，Web 架构提供了两种方式来解决其在 Web 上的访问问题：一种是 Hash URIs，另一种是 303 URIs。

Hash URIs 方式是采用带有“#”分隔符的 URI 地址命名非信息资源，如将元数据本体中定义的抽象概念 Book 命名为<http://hostname/ontology/core#Book>。当一个非信息资源的 Hash URI 地址

① Berners-Lee, T. Linked data – design issues [EB/OL]. [2010-07-01].<http://www.w3.org/DesignIssues/LinkedData.html>.

② Berners-Lee T. Linked data -design issues [EB/OL] [2010-07-01].<http://www.w3.org/DesignIssues/LinkedData.html>

③ 在 Web 开发中，一个 mashup 是一个网页或应用，它使用和合并来自两个或多个数据源中的数据、表示或功能从而生成新的服务。

④ Bizer, C., Cyganiak, R., Heath, T. How to Publish Linked Data on the Web [EB/OL]. [2010-07-01].<http://www4.wiwi.fu-berlin.de/bizer/pub/LinkedDataTutorial/>.

被浏览器请求时,在向服务器发出请求之前,HTTP 协议会自动将 URI 地址中“#”符号之后的部分剥离掉,因此向服务器真正请求的是一个能够被 HTTP 协议解引用的信息资源的 URI 地址,即元数据本体的 OWL 文档地址<http://hostname/ontology/core#>,然后服务器将该信息资源的一个表示(如 RDF/XML 文档)返回给浏览器,这个表示包含了对被请求的非信息资源(即 Book 概念)的描述。Hash URIs 访问方式适用于小型的 RDF 词表(如本体),浏览器可以很快显示整个词表文档,而且因为文档长度较小易于浏览,但是对于含有大量 RDF 三元组的实例文档则不适用。

303 URIs 方式采用带有“/”分隔符的 Slash URI 地址命名非信息资源,如将一本图书命名为<http://hostname/document/book/isbn9787301149034>。当一个非信息资源的 URI 地址被浏览器请求时,服务器根据客户端浏览器的类型将其重定向到描述它的一个信息资源的 URI 地址:如果是 HTML 浏览器,服务器倾向发送 HTML 文档,如<http://hostname/document/book/isbn9787301149034.html>;如果是 RDF 浏览器,服务器倾向发送 RDF/XML 文档,如<http://hostname/document/book/isbn9787301149034.rdf>。然后浏览器再向服务器请求这个新的 URI 地址,服务器返回 HTML 或 RDF/XML 文档,它提供了对被请求的非信息资源的描述。因此,对于一个非信息资源,303 URIs 方式需要命名三个相关的 URI 地址:① 资源本身的 URI 地址;② 资源元数据的 RDF/XML 表示;③ 资源元数据的 HTML 表示。但是,采用 303 重定向访问的一个主要缺点是需要两次 HTTP 请求才能获取一个非信息资源的描述,因此会造成访问延迟。

8.5.3 关联数据中资源命名原则

在关联数据背景下,所有资源对象的命名都需遵循以下原则:

(1) 采用 HTTP URIs 地址命名任何事物,包括信息资源对象、个人/组织机构/团体、时间日期、地点以及受控词表中的概念和关系等。

(2) 在一个你能够控制的 HTTP 命名空间里定义 URIs 地址,而不是其他人的命名空间。因为在一个能够控制的命名空间,你可以真正使 URIs 地址被解引用。譬如国家图书馆的所有资源均使用国家图书馆的命名空间<www.nlc.gov.cn>进行命名。

(3) 最好采用比较短小、易于记忆的 URIs 地址。

(4) 最好使用稳定、持久的 URIs 地址,因为改变 URI 地址将会使已经建立的数据间的链接被破坏。因此在命名任何事物时,对选择的 URIs 地址的命名规则都要经过慎重考虑,避免以后更改。

(5) 选择的 URIs 地址通常要受到技术环境的限制。譬如,如果服务器不能使用默认的 80 端口作为 HTTP 协议的端口号,则必须在主机名后加上端口号,如<http://www/nlc.gov.cn:2020>,当然可以通过对 Web 服务器进行设置将 URIs 地址重写成比较简单的形式。

(6) 对于一个非信息资源,通常要命名三个相关的 URI 地址:

- 资源本身的 URI 地址(建议无任何扩展名);
- 资源元数据的 RDF/XML 表示(建议以 rdf 作为扩展名)
- 资源元数据的 HTML 表示(建议以 html 作为扩展名);

譬如一本图书的上述三个 URI 地址分别为:

- http://hostname/resource/004106310
- http://hostname/resource/004106310.rdf
- http://hostname/resource/004106310.html

另一种表示方法是:

- http://hostname/resource/004106310
- http://hostname/data/004106310(表示 RDF 文档)

- <http://hostname/page/004106310>（表示 HTML 文档）

(7) 对于 RDF 词表文档（如本体），因为所含成员（即类和属性）数量较少，建议采用带有“#”的 Hash URI 地址命名词表成员。“#”号之前的部分是词表文档的 URI 地址，之后部分是词表成员的本地标识符。因为访问 Hash URI 地址时不需重定向，浏览器可以很快地显示包含词表成员的整个词表文档，而且因为文档长度较小，易于浏览。

(8) RDF 实例数据，建议采用带斜线的 Slash URI 地址命名，通过 303 重定向方式进行访问。

(9) 通常在 URIs 地址中要包含某种主键值来保证每个 URI 地址的唯一性，譬如采用图书馆中的记录号作为文献资源对象的本地标识符，如 <http://hostname/resource/004106310>，或者采用 ISSN 和 ISBN 号作为本地标识符。

8.5.4 关联数据发布方法

目前关联数据的发布主要有以下五种方式^①：

(1) 以静态 RDF/XML 文件发布关联数据：利用 Web 服务器（如 Apache HTTP 服务器）的 URL 重写功能和 HTTP 内容协商机制将非信息资源（即实体对象或抽象概念）的 URL 地址重定向到描述它的信息资源（如 HTML 或 RDF/XML 文档）的 URI 地址，HTML 或 RDF/XML 文档采用离线的方式预先手工或自动创建。这种方式通常用于发布小型的 RDF 词表，但是对于大数据量却并不适用，因为这需要预先生成大量的 HTML 或 RDF/XML 文档。

(2) 采用服务器端脚本发布关联数据：通过服务器端脚本（如 PHP）基于后台的关系型数据动态地生成 HTML 或 RDF/XML 文档（需通过 ARC^②类库），或者通过 SPARQL 终端直接从 RDF 存储器中获取 RDF 数据，然后利用服务器端脚本或者脚本与 Apache 服务器的 URL 重写功能相结合以实现非信息资源 URL 地址到相应的信息资源表示（即 HTML 或 RDF/XML 文档）的重定向。

(3) 以 RDFa 格式发布关联数据：采用 RDFa 格式^③将 RDF 三元组内嵌在 XHTML 网页中，然后利用 Web 服务器的重定向功能将非信息资源的 URL 地址重定向到描述它的 XHTML 网页（针对 HTML 浏览器）或者重定向到从 XHTML 网页中提取出的 RDF/XML 文档（针对 RDF 浏览器）。

(4) 从 RDF 存储器发布关联数据：利用 RDF 三元组存储器（如 Jena、Sesame、AllegroGraph^④等）直接存储 RDF 数据，这些存储器通常带有一个 SPARQL 终端（如 Jena 的 Fuseki），能够支持基于 Web 的 SPARQL 查询和结果显示，但是无法在浏览器中访问非信息资源的 URI 地址。此时可在 RDF 存储器的 SPARQL 终端的前端放置一个关联数据界面（如 Pubby^⑤），将不可解引用的 URI 地址转换为能够被 HTTP 协议解引用的，实现关联数据显示。

(5) 从关系型数据库发布关联数据：利用现成的工具将存储在关系型数据库中的关系型数据直接发布为关联数据。最广泛使用的工具是 D2R 服务器^⑥，它能够帮助用户在关系型数据库结构和 RDF 术语间建立映射，对关系型数据生成一个关联数据视图，支持 RDF 浏览器对关系型数据的关联数据化显示和 SPARQL 终端对关系型数据的查询。其他类似的工具还包括小型的开源工具 Triplify^⑦和商业软件 OpenLink Virtuoso^①。

① Bizer, C., Cyganiak, R., Heath, T. How to Publish Linked Data on the Web [EB/OL]. [2010-07-01]. <http://www4.wiwi.fu-berlin.de/bizer/pub/LinkedDataTutorial/>.

② ARC 是一个支持 RDF 和 SPARQL 的开源 PHP 工具包。

③ RDFa 是 W3C 的一个推荐标准，它通过扩展 XHTML 的几个属性使 RDF 三元组能够内嵌在 XHTML 网页中，RDF 三元组可通过 RDFa Distiller & Parser 从 XHTML 网页中自动提取出来。

④ AllegroGraph 是由 Franz Inc. 公司开发的一个商业化的 RDF 数据库和应用框架。

⑤ Pubby.[EB/OL].[2010-07-01].<http://wifo5-03.informatik.uni-mannheim.de/pubby/>

⑥ D2R Server: Accessing databases with SPARQL and as Linked Data.[EB/OL].[2010-07-01]. <http://d2rq.org/d2r-server>

⑦ 详见：<http://triplify.org/Overview>

(6) 通过包装已有的应用或 Web APIs 发布关联数据：通过构建关联数据包装器将目前已有的多个应用或 Web APIs 包装到一个语义网访问界面中，使得原本需要用户通过各个不同的应用或 API 访问的数据能够通过这个统一的界面以关联数据的形式进行访问。包装器的作用是将用户对 URI 地址的请求转换成对各个应用或 API 的查询，然后将各自返回的查询结果进行集成并转换成 RDF 格式发送给用户。

8.5.5 关联数据应用实例

本节以图书《数字图书馆的知识组织系统：从理论到实践》为例，说明关联数据的发布和浏览过程。

该图书的 URI 标识符是<<http://www.nlc.gov.cn/resource/004106310>>，对于采用 Slash URI 地址命名的非信息资源需采用 303 重定向方式进行访问，参见图 8.5 和图 8.6。采用重定向方式访问的一个缺点是会有一些延迟，一个解决策略是在资源的 URI 标识符后添加“#this”将 Slash URI 地址转换成 Hash URI，然后采用 Hash URI 方式进行访问。如图 8.15 所示，当浏览器向服务器发出对 URI 地址<<http://www.nlc.gov.cn/resource/004106310#this>>的请求之前，HTTP 协议自动将“#this”从 URI 地址中剥离掉，向服务器请求的实际地址是<<http://www.nlc.gov.cn/resource/004106310>>。该地址代表的是一个信息资源，具有多种表示形式（如 HTML、RDF/XML 和 Text/N3），服务器通过内容协商机制选择合适的表示发送给客户端浏览器。

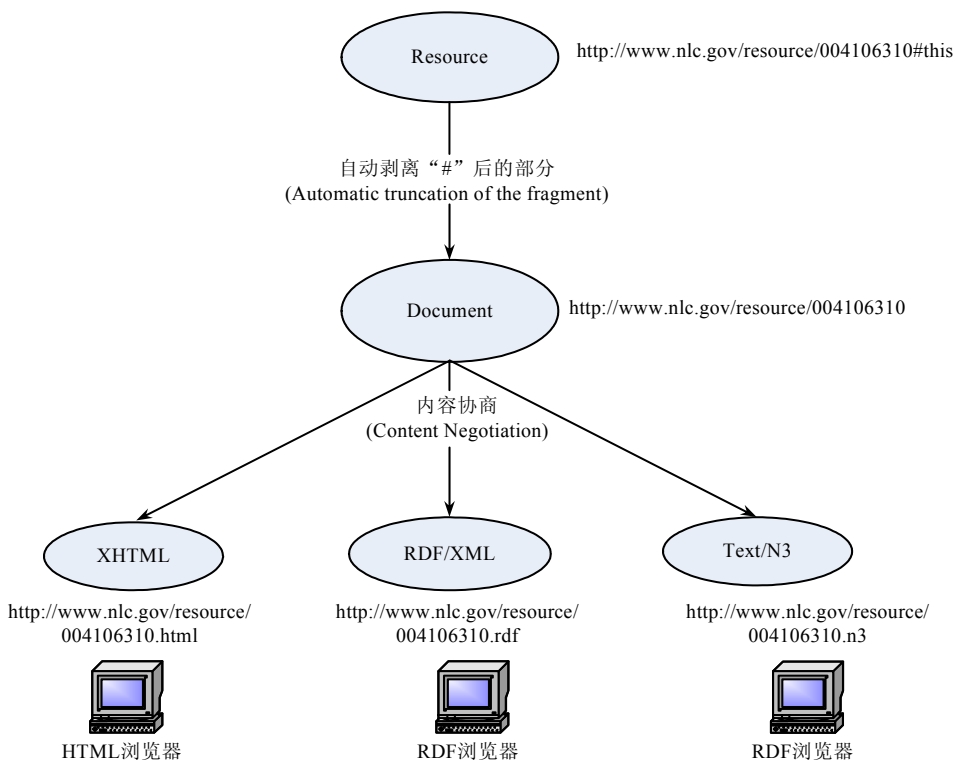


图 8.15 Hash URI

① 详见：<http://virtuoso.openlinksw.com/>

下面详细说明使用内嵌在 Firefox 中的 Tabulator RDF 浏览器^①浏览资源<<http://www.nlc.gov.cn/resource/004106310#this>>及其相关资源的整个过程。

(1) 输入 URI 地址“<http://www.nlc.gov.cn/resource/004106310#this>”，浏览器显示该图书的 RDF 元数据表示，如图 8.16 所示。

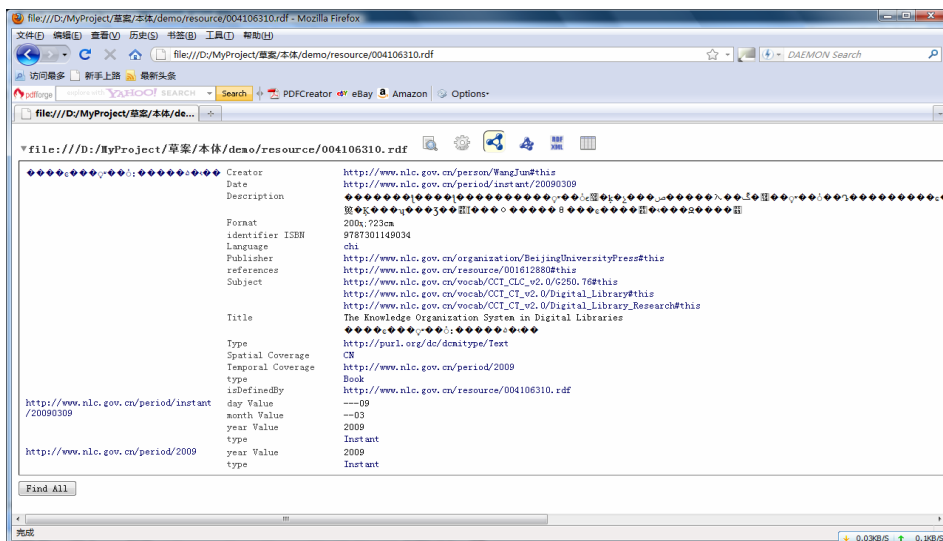


图 8.16 图书的 RDF 元数据描述 (004106310.rdf)

(2) 点击“<http://www.nlc.gov.cn/person/WangJun>”，链接到作者的 RDF 元数据表示，如图 8.17 所示。

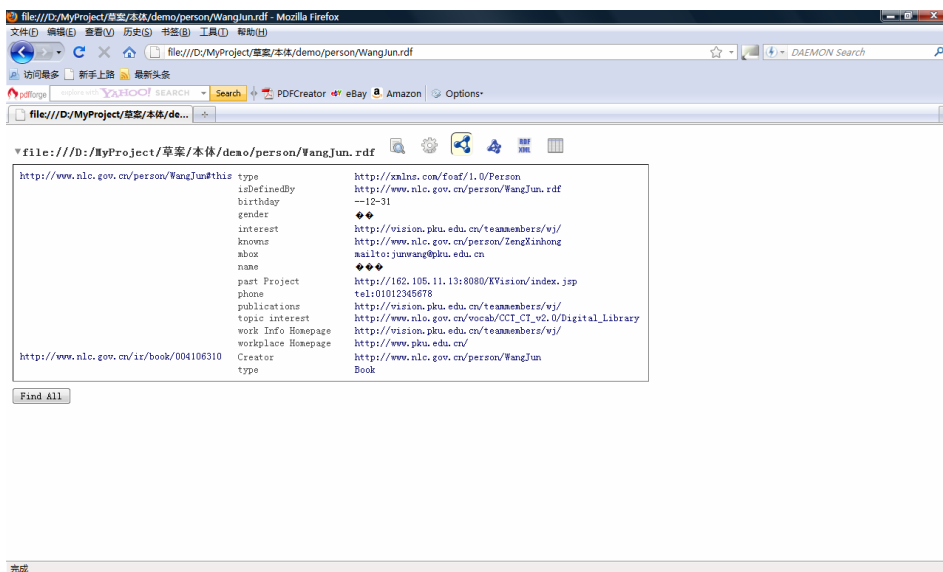


图 8.17 作者的 RDF 元数据描述 (WangJun.rdf)

^① Tabulator 是一个数据浏览和编辑器，提供了在 Web 上浏览 RDF 数据的途径，可以作为 Firefox 浏览器的扩展附件与该浏览器一起使用，见 <http://www.w3.org/2005/ajar/tab>。(目前还无法支持汉字显示。)

(3) 点击“http://www.nlc.gov.cn/vocab/CCT_CT_v2.0/Digital_Library”，链接到“汉语主题词表”中对“数字图书馆”这一概念的 RDF 描述，如图 8.18 所示。

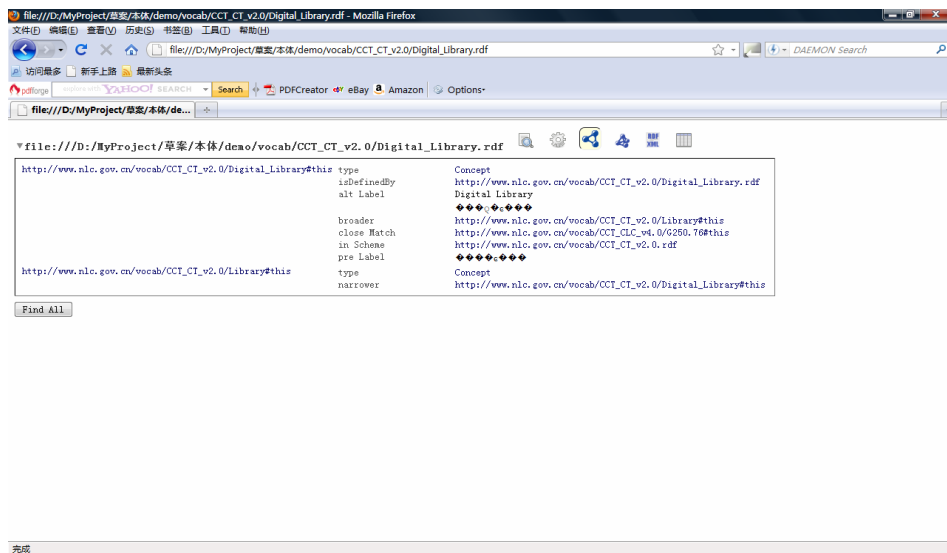


图 8.18 “汉语主题词表中”对“数字图书馆”这一概念的 RDF 描述 (Digital_Library.rdf)

(4) 点击“http://www.nlc.gov.cn/vocab/CCT_CT_v2.0/Library”，链接到“汉语主题词表”中对“数字图书馆”的上位概念“图书馆”的 RDF 描述，如图 8.19 所示。

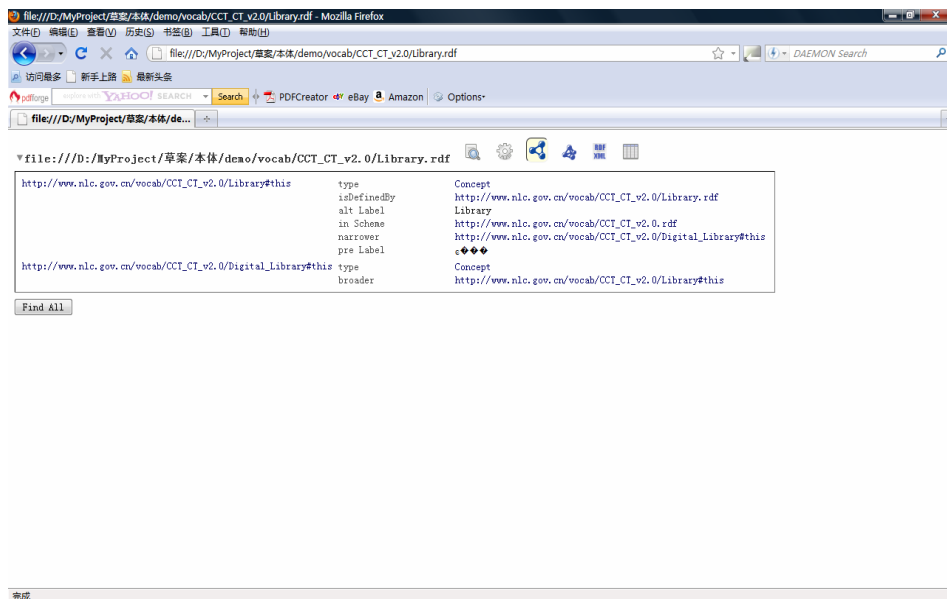


图 8.19 “汉语主题词表中”对“图书馆”这一概念的 RDF 描述 (Library.rdf)

(5) 回到所检索的图书的元数据页面，点击“http://www.nlc.gov.cn/vocab/CCT_CLC_v4.0/G250.76”，链接到“中国图书馆分类法”中对分类号“G250.76”的 RDF 描述，如图 8.20 所示。

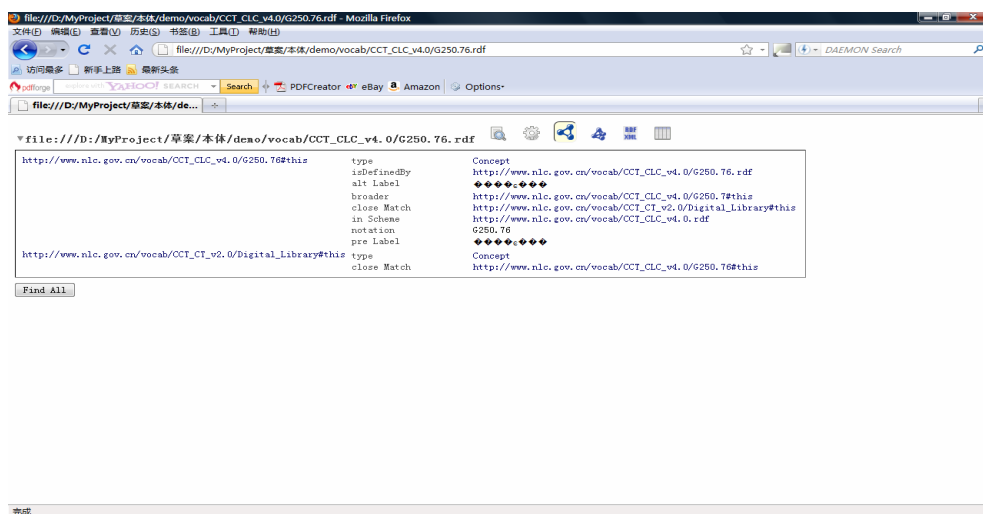


图 8.20 “中国图书馆分类法”中对“G250.76”分类号的 RDF 描述

(6) 回到被检索的图书的元数据页，点击“http://www.nlc.gov.cn/onto/core_v1.0.owl#Book”，链接到 NLOC 核心元数据本体的 OWL 文档，其中包含了对概念“Book”的 RDF 描述，如图 8.21 所示。

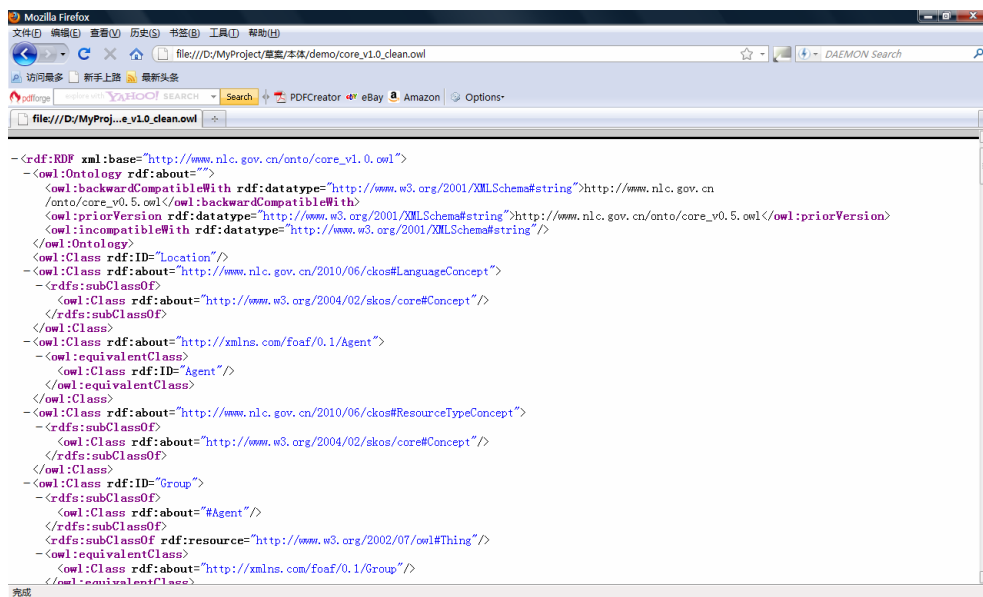


图 8.21 NLOC 核心元数据本体的 OWL 文档的 RDF/XML 表示



本章小结

本章论述了语义网环境下的信息组织。首先介绍了语义网中的信息描述与表示格式 RDF 数据模型、语义网中的信息建模方式本体及其构建方法、以及语义网中的知识组织系统描述

语言 SKOS, 然后举例说明在语义网环境下如何采用 OWL 本体对领域知识进行建模, 如何采用 RDF 语言基于本体对信息进行语义化的描述, 如何将描述好的信息在网络上发布为可访问的关联数据。



问题讨论

1. 语义网与当前 Web 有何区别与联系?
2. 构建本体的基本方法有哪些?
3. 基于语义网的信息组织有何特点与优点?
4. 什么是术语注册与术语服务?
5. 关联数据的发布方式有哪些?



第9章

不同环境下的 信息组织评价

内容提要

信息体经过著录形成众多款目或记录，信息组织工作实质上就是对这些款目或记录用词汇或分类号标引后形成各种检索工具或检索系统。这些检索工具或检索系统按照载体形式可以划分为印刷型、多媒体型、网络型检索工具或系统为主的各种类型。本章将提出这几种信息组织方式的评价标准，列举出各类型信息组织的成果实例，并对其价值、优劣等就行评价。由于不同类型和载体的检索工具或系统有不同的特点、方式、目的等，因此其评价指标也有所差异。

本章重点

- 各类型资源信息组织的特点；
- 各类型资源信息组织的评价标准；
- 各类型中有代表性的信息组织实例评价。



9.1 非网络环境下的信息组织评价

9.1.1 印刷型文献的信息组织评价

印刷型文献是以纸张为存储介质，以印泥、铜印、胶印、油印、铅印、静电复印等为记录手段而产生出来的一种传统的文献形式，如图书、期刊、会议文献、科技报告等。其优点是便于阅读和流传，不受时间、地点、设备等条件的限制。其缺点是存储密度低，篇幅、体积庞大，容易破损，占据储藏空间过多，难以实现自动检索。

印刷型文献的信息组织主要是通过一些传统的信息组织技术如文摘、综述、目录和索引来进行组织的，从而形成一系列印刷型检索工具，主要有以下七大类。

（1）目录、索引、文摘。

目录也称书目，是著录一批相关图书或其他类型出版物，并按一定次序编排而成的一种检索工具。索引是记录一批或一种图书、报刊等所载的文章篇名、著者、主题、人名、地名、名词术语等，并标明出处，按一定排检方法组织起来的一种检索工具。文摘是以提供文献内容梗概为目的，不加评论和补充解释，简明、确切地记述文献重要内容的短文。汇集大量文献的文摘，并配上相应的文献题录，按一定的方法编排而成的检索工具，称为文摘型检索工具，简称文摘。

（2）百科全书。

百科全书是参考工具书之王，概述了人类一切门类或某一门类知识的完备工具书，是知识的总汇。百科全书一般按条目（词条）字顺编排，另附有相应的索引，供迅速查检。

（3）年鉴。

年鉴是按年度系统汇集一定范围内的重大事件、新进展、新知识和新资料，供读者查阅的工具书。它按年度连续出版，可用来查阅特定领域在当年发生的事件、进展、成果、活动、会议、人物、机构、统计资料、重要文件或文献等方面的信息。

（4）手册、名录。

手册是汇集经常需要查考的文献、资料、信息及有关专业知识的工具书。名录则提供的是一有关专名、人名、地名、机构名等的简明信息。

（5）词典、字典。

词典是最常用的一类工具书，分为语言性词典和知识性词典两种。

（6）表谱、图录。

表谱是采用图表、谱系形式编写的工具书，大多按时间顺序编排，主要用于查检时间、历史事件、人物信息等。

（7）类书、政书。

类书是我国古代通过摘录、汇辑多种文献中的原文、按内容性质分门别类地编排组织，以供寻检和征引的工具书，是我国兼有“百科全书”和“资料汇编”性质的工具书。政书则是我国古代的一种专史著作，它分门别类地汇集历代或某一朝代的政治、经济、军事、文化制度等资料，具有资料汇编性质和便于检索的特点，是查找我国古代典章制度的重要工具书。

随着信息技术及网络的发展和普及，曾在文献检索中发挥重要作用的目录、索引和文摘大多已以数字化和网络化的形式出版，不再属于传统印刷型检索工具，如《人大复印报刊资料索引》、《全国报刊索引》等。目前常用的印刷型检索工具中，词典和百科全书的编制体例丰富、检索系统完备，年鉴、手册及名录、表谱和图谱则相对简单。因此，本节选取词典和百科全书为例来阐述印刷型文献信息组织的标准，但对于不同类型的印刷型检索工具而言，

其具体信息组织评价标准是有区别的。

1. 词典信息组织的实例与评价

词典是汇集语言里的词语,主要解释词语的概念、意义及其用法,并按一定方式编排,以便查阅的工具书。它具有索解性、简明性和规范性的特点,是一种非常重要的工具书。

1) 词典的组织结构

由于索解性的特点,与普通图书相比,词典的组织结构更为复杂。它一般包括三个部分:正文、辅助说明和目录索引、相关附录。

(1) 正文。

正文即解释词语的内容部分,它是词典的主体,包括3个部分:排列、注音、释义和例证。排列是指词的编排方法,词典的编排方法主要有部首法、笔画法、笔形法、四角号码法、笔形号码法、声母法、韵母法、注音字母法、汉语拼音字母法、分类法这10种形式。注音即词的注音方法,主要有读若法(读相似的音,如《说文解字》的“读若某”)、直音法(用同音字注音)、反切法(用两个汉字拼切替另一个汉字注明读音)、注音字母法(用37个注音字母再加四声注音)、汉语拼音字母法(用26个汉语拼音字母再加四声注音)5种方法。释义是正文最主要的内容,词典正是通过释义而提供知识的。释义主要有以下4种方法:义训(即从训诂研究词义、采用同义词对释互训)、形训(即从字形推求字义)、声训(即从字音发现字义)、定义(即用词组、语句给词下定义,确定词义的内涵)。例证是词典释义的重要辅助手段,为词的释义提供书证和举例。

(2) 辅助说明和目录索引。

辅助说明包括序言、凡例等,一般在正文之前,通过它可了解编制目的、使用对象、取材范围、编排方法等。它是读者了解词典概况和使用方法的最好来源。目录一般在正文之前,用来提供整个词典的体系结构。索引一般在正文之后,为读者提供多种检索途径,二者的目的均为便于检索。

(3) 相关附录。

词典最后常附载有价值的参考资料,或补充正文、或别有用途另供参考。例如《现代汉语词典》的附录就有我国历代纪元表、计量单位表、汉字偏旁名称表、汉语拼音方案、元素周期表和中国地图;《辞海》的附录也多达14种。

2) 词典信息组织的评价指标

词典作为一种供读者查阅以释疑解惑的工具书,其信息组织的评价应以是否能够提供便捷的检索功能为中心,具体有以下几个评价指标。

(1) 编排顺序。

词典的编排顺序直接反映词条款目组织的顺序,该顺序一旦确定也就确定了整个词典的信息组织的基本方式。正因为它的重要性,才在编制各类词典时产生了多种编排方法。例如,新中国成立以来,随着汉语拼音方案的推出和规范,现代词典的编排顺序常用的是汉语拼音字母法,拼音文字按字母顺序排列比较易查,如《现代汉语词典》、《新华词典》等。除此之外,《说文解字》这些大部头的资料型的词典采用部首法进行排序。《广韵》、《集韵》、《佩文诗韵》这类以“韵”为特色的词典采用的是韵母法。《尔雅》还以分类法作为词典的编排方法,包括释诂、释言、释山、释水等19篇,便于读者从类的角度查寻相关词条。还有人尝试按词族原则编排,优点是能显示同族词之间的结构语义联系,但终因查检不便而遭到读者的反对。此外,有的词典词条中收录了大量的成语,但并未按打头词首字母顺序排列,因而也给读者带来查检上的不便。无论采取哪种编排顺序,不同类型的词典都必须根据自身特性和读者需求的特点来选择最适合的编排方式。

（2）目录索引。

词典的目录提供给读者的是词典的整体信息组织结构，包括总目、凡例、各种索引、词典正文、辅助说明和附录等，用户可以借助目录找到相关部分。

如上文所述，词典的编排顺序是唯一的，且一旦确定就不可更改，但读者检索需求具有多元性，单一的检索途径很难满足读者需求。另一方面，单一的排检次序也不利于全方位地揭示词条款目间的关系。词典的索引正是解决这个问题的有效途径，它不仅是增加词典检索途径的最有效手段，同时也使词条款目的多途径组织成为可能。因此，词典索引编制的成功与否是词典信息组织的关键之一。一般来说，词典都会采用多种索引来对编排次序进行补充，如《辞海》除每卷书前有《辞海部首表》之外，其索引还包括《笔画索引》、《汉语拼音索引》、《四角号码索引》、《词目外文索引》，便于读者从部首、笔画字形、汉语拼音和四角号码等多个角度进行检索，查找所需词的信息。虽然索引非常有用，如果不切实际地编制过多的索引，则不仅不会为读者提供方便，还会造成编制成本的激增。词典索引的评价标准与图书索引相比，更多地强调规范性、易检性和简洁性。因此，索引的编制设计应当以读者为中心，以提供便捷的检索途径为宗旨。

（3）参见系统。

参见（或称引见）是词典中的一个重要环节，一部词典是一个统一的有机整体，其内部纵横交错，互相疏通，互相补充，互相印证，参见系统揭示的是词条款目间的相关关系，读者可以通过参见系统从一个词条方便地转入相关词条。它的严密与否也关系着词典信息组织的成败。

参见系统是否严密可从以下5个方面来衡量。

① 参见是否有遗漏。例如，词目A见词目B，但词典中却无B词条；又如，词目A见词目B，而词目B又见词目A，结果两头落空。

② 参见是否有重复。参见是词典编纂中既节省篇幅又提供相关知识的重要手段，本应采用参见办法解决问题，但却重复释义和举例，不仅浪费了篇幅，又未给读者带来新的信息，甚至有时前后还会出现释义不一致的情形，而释义的一致性恰恰是词典释义的一个重要原则。

③ 参见是否沟通了相关条目间的联系。语言是个完整体系，但词典按单词立目进行解说，客观上破坏了语言词汇的体系性，完善的参见系统可以在一定程度上弥补这种缺陷，给读者提供相应的语文知识和百科知识，如同义近义关系、上下义关系、派生关系及词的各种变体形式等。

④ 参见是否有必要的注释。如上所述，参见系统提供相关知识，而相关不等于相同，在使用参见词时要注意使用相关注释，以指出修辞、语体、地域等差异。

⑤ 参见的交叉条目的互见是否合理。交叉条目在汉语词典中往往分属于不同的领头字，这就产生了检索性问题。在互见过程中，应将各相关交叉条目的参见指向同一条目，以免造成混乱。

（4）体例说明。

体例，即著作的体裁、凡例，是词典信息组织的辅助手段。词典中使用代码的最大优点是节省篇幅，使词典信息组织变得简捷明了，便于读者使用。但代码若比较烦琐，读者若望而生畏，掌握不熟练，反而平添许多查检上的麻烦。因此，词典体例安排是否合理、是否科学关系到读者的检索使用效率。

体例反映了一部词典的面貌，体例混乱必然会降低词典的质量。同一部词典的体例必须绝对统一，这既便于查检，同时免去读者在查检中可能产生的困惑。体例说明中的释文模式要恰当，根据不同的词典规模和性质，要选择恰当的释文模式，既不能一味追求详尽而显得啰唆，也不能过于简练而不便于读者理解。

3) 词典实例《现代汉语词典》的评价

《现代汉语词典》(以下简称《现汉》)出版以来,为推广普通话、促进汉语规范化工作做出了重要贡献,在我国文化教育和科学研究事业中发挥了巨大作用,受到读者的欢迎和社会的重视,曾荣获国家图书奖、国家辞书奖、中国社会科学院优秀科研成果奖、吴玉章人文社会科学奖,在海内外享有盛誉。《现汉》由专业权威的编写机构中国社会科学院语言研究所编撰,是现代汉语规范的权威之作,代表了我国当今词典编撰的最高水平。《现汉》1978年正式出版后,曾于1983年、1996年、2002年、2005年出版过修订本或增补本,最新的版本为2012年出版的第6版。除了增收新词、删减旧词及修改释义和例句外,编者还复查了词类标注,在保持原有词类标注体系的基础上,对少数词的词类标注做了修订,根据有关标准和新的研究成果对检字表和附录做了修订。

在编排顺序方面,《现汉》大体上采用汉语拼音字母次序排列,形同而音、义不同的分立条目。具体来说,单字条目按拼音字母次序排列;同音字按笔画顺序排列,笔画少的在前,多的在后;笔画相同的,按起笔笔形横、竖、撇、点、折的顺序排列。单字母条目之下所列的多字条目不止一条的,依第二字的拼音字母次序排列;第二字相同的,依第三字排列;以下类推。轻声字一般紧挨在同形的轻声字后面。《现汉》作为一部推广普通话的权威工具书来说,采用这种编排顺序是恰当的。

在目录索引方面,《现汉》的目录包括了凡例、各种查检表(《音节表》、《新旧字形对照表》、《部首检字表》)、词典正文、附录四个部分。供读者查检使用的索引包括《音节表》、《部首检字表》,读者可通过字音和字形两种方式进行检索。《部首检字表》又分为《部首目录》、《检字表》和《难检字笔画索引》三部分,前两者的结合使用可检索到逐字层次,为方便读者查检,有些字还分散在几个部首内,如“思”字在“田”部和“心”部都能查到。《难检字笔画索引》则供读者查找一些不易通过区分部首偏旁来查检的字,如“〇”、“乐”、“臧”等。《现汉》的索引虽然不如《辞海》等大部头的词典丰富,但由于其广泛的适用性和普及性,它提供的这些简练但实用的检索途径也能够满足读者日常查检需求。

《现汉》的参见系统按字词音、形、义区分,可分为三种:一是字、词义相关但音、形不同,如“金牛座”参看571页“黄道十二宫”;二是字、词形同但音、义不同,如“闷(mēn)”另见887页mèn;三是字、词义同但形不同,如“炎帝”另见1497页“炎黄”。对于这三种词义不同的情况,《现汉》的相关参见中均标识了参见到的页码。另外,对于同义词,《现汉》在词条中标注“也说…”、“也叫…”、“也作…”、“注意…”等表示同义词。这个参见系统由于采用了相关词页码标识和同义词互见相区分的方式,因此参见不易遗漏。对于同义词,《现汉》没有将所有的同义词逐一注明含义,而是将同义词都指向同义词群中的某一个词条,这既避免了参见的重复和交叉条目参见的混乱,又保持了释义的一致性,如“单车”、“脚踏车”都标注见“自行车”。在参见注释方面,《现汉》使用相关注释来指出相关词在使用时的修辞、语体、地域等方面的差异,如“潺潺”注释为拟声,“藏猫儿”注释为<口>，“寿司”注释为[日]。但《现汉》的参见系统也存在一定的缺陷,在表示词间关系方面,它没有指明具体的词间关系如近义、派生还是包含关系等,只是采用各种参见的标名如“见”、“另见”、“参看”、“同”、“也作”等进行区分,但在体例说明部分也没有说明这些标名表示的是何种关系。

在体例方面,经过全面修订的《现汉》第6版较之以前的版本更加科学严谨,但在一些体例上还存在有待完善的地方。例如,姓氏义项的处理方式不统一;“也叫”、“也说”后非常用词形的处理方式及“全称”、“简称”的处理方式不一致,体例不统一等。

2. 百科全书信息组织的实例与评价

百科全书是系统概述人类各个知识门类或某一知识门类的基本知识,按辞典形式编排的

大型工具书。可以说，百科全书包括了人类全部知识体系，具有概述性、完备性和权威性等特点。百科全书除了教育功能外，还是一种非常重要的工具书，这也决定了百科全书必须具有检索性。而由于百科全书部头大、知识覆盖领域广，它的检索性与其他工具书相比又有其特殊之处，因此它的编撰体例和检索系统也最完备，其信息组织系统是所有纸质文献中最复杂的。

1) 百科全书的编排体例

(1) 编排方式。

百科全书有按知识分类和按条目字顺编排两种方式。西方现代百科全书大多按字母顺序编排，又分为逐词排序（Word by Word）和逐字排序（Letter by Letter）两种。如果条目的单词相似，内容则按人名、地名、物名的先后顺序排。同姓名的人物应该根据人名后的注释做出判断。按字顺排列的百科全书多通过参照系统反映条目之间的学科关系。这一方法易于检索，但打乱了整个知识领域的系统性，不利于学科总体研究。

正文条目按学科体系排序这一排列方法使知识体系清楚、明确，但分类方法不易掌握，影响读者对百科全书的使用，因此卷末会辅以一个详尽的字顺索引，并且该字顺索引质量必须高。

(2) 条目编撰。

百科全书不同于词典等其他的工具书。词典等是对词语条目进行释义，而百科全书用条目的形式介绍知识。因此百科全书对条目的阐释也比其他工具书详细，经常需要分成许多的小标题。

在条目编撰方面，注重教育功能的百科全书以大条目为主，全面系统地介绍某学科领域的知识，保持知识的系统性和完整性，但必须以完善的分析索引来弥补不便于查找某一具体材料的缺陷。注重检索功能的则采用小条目的编纂思想，把知识单元划分得很小，这对检索具体的材料比较方便，但不能使读者得到系统的知识，而必须辅之以具有一定水平的内容分类表。一般来说，百科全书条目的构成部分有：条头、释文、插图、参考书目。

(3) 参见系统。

参见系统的使用是把相互关联的条目和概念联系起来，向读者指示与所查内容有关的条目，帮助读者获得一个主题更完整和更系统的知识。它往往隐藏在内容体系之中，提供与本条有关的其他条目，给读者排检带来巨大的方便，加强了百科全书的工具书作用。同时，它又将被分离但实际又有关联的条目联系起来，加强了百科全书的教育作用。百科全书的教育作用和工具书作用是相互矛盾的，而参见系统恰恰可以在一定程度上解决这个矛盾。因此，百科全书的参见系统应用是非常广泛的，按处理方式分可分为随释文分散参见、分段集中参见、条末集中参见和几种方式结合运用这四种参见系统。

参见系统包括参见条和条目释文内的参见。参见条是同一概念的重要异称，或者它的内容已另设专条。例如美国著名作家 Mark Twain（马克·吐温），他的原名是 Samuel Langhorne Clemens（塞姆·朗赫恩·克列门斯），读者无论用笔名或用原名都能查检到。

(4) 索引和其他检索系统。

百科全书索引是把百科全书正文中可用做标识的各种主题、专有名词、符号或其他对象的名称分析、摘录出来，注明页码，按一定次序编排起来，以便进行相应途径检索的辅助性工具。它对所述内容多层次、多角度地揭示，为不同需要、不同目的的读者提供多条检索途径。它作为一种微观检索的标识，只是指出哪里有所需资料，而并不是提供所要获得的资料本身，它是加强百科全书查询功能的最主要的措施。

百科全书的主题索引一般分为简式和复式两种，而索引一般按照字顺排列。还有的把索引变成小百科词典，款目逐条注出其内容的简明解释，独立成书，加上卷次和页码，又成为

其百科全书正文的索引,如第15版《新不列颠百科全书》(New Encyclopedia Britannica 15thed)。

百科全书后的附录内容一般都比较丰富,如《兰登书屋百科全书》(The Random House Encyclopedia)的附录有“大事年表”、“全世界的旗帜”、“图片一览表”、“艺术一览表”等。这些附录虽与正文没有直接联系,但都可作为独立的参考书使用,从而为读者提供新的检索手段。

2) 百科全书信息组织的评价标准

由于百科全书编排体例的复杂性,其信息组织评价标准与词典相比,也有综合性、交叉性的特点。

(1) 条目编排方式是否合理。

无论是按字顺排列还是按知识分类排列,都不存在孰优孰劣的问题,重要的是是否采用多种索引对之进行补充。

正文按字顺排列的百科全书是否辅以各种内容主题索引、参见系统和表现全书知识体系和引导系统自学的条目分类目录、学习指南进行补充?正文按学科体系分类的百科全书是否辅以高质量的详尽的条目字顺式索引以便读者排检?

(2) 条目的选择是否合理。

条目是百科全书的基本查阅单元,其条目的选择与撰写恰当与否直接影响着百科全书的信息组织质量。

在条目层级分工方面,作为知识具有层次结构的反映,百科全书的条目是有着层次划分的。这里所说的层级分工是指在选条中通过分别安排上、中层综述性条目和下层具体人、地、事、物条目,去满足读者不同层次的检索需求,以功能有别的不同层级条目共同完成介绍知识的任务。在百科全书编纂中,大条目主义和小条目主义均不要求层级分工,前者提倡把下层小主题包容到上层大条目中去,主要用上层大条目包揽对知识的介绍;后者主张把下层小主题从上层大主题中尽量分解出来,形成大量小条目,同时把上层大主题进行类辞典式条目处理,从而使上下层条目篇幅接近,在功能上无明显差别。但在读者实际使用中,条目层级分工是非常必要的,单一的条目层级不能满足读者需求。具体来说,读者有时需要检索人、地、事、物一类有限主题,如果把这些小主题都包容到上层大条目之中,依靠内容索引解决它们的检索问题,实际上读者得不到完全满意的答案。这是由于必须服从大主题本身的介绍,小主题的内容往往支离破碎、残缺不全,甚至无法保证内容要素的齐备。为了满足读者的检索需求,让百科全书充分具备工具书性能,就应当把检索率高的小主题设为条目,以保证它们内容的完整及检索的便利。而另一方面,在另一些情况下读者需要查检范围较大的主题,如果在选条时把大、中主题安排为类词典式条目,缩小它们的篇幅,规定仅做有限的内容展开,那么读者就可能无法从中查到其所需要了解的某一领域的全貌及相关规律性。只有在选条中设置进行展开介绍的大、中型综述性条目,才能充分体现出百科全书的特点,使百科全书具备辞典所没有的功能。百科全书发展的历史中一直存在着“大条目主义”和“小条目主义”之争,也证实了条目层级分工的合理性。体现在编撰实践上,大条目的知识系统性好,但不便具体问题的寻检;小条目便于特定问题的查检,却又削弱了知识的系统性。鉴于它们各自的优劣利弊,条目层级分工是百科全书选条的一项重要原则,具体怎样分工,各百科全书应该本着自身特点而有不同的具体形式。

在百科全书条目的选择标名方面,条目标名与检索关系极大。条目标名应当符合准确性原则、通用性原则、名词性原则和简要性原则。准确性原则是指条名必须准确标识条目主题,不应当是文不对题或似是而非、不能语义含混让人捉摸不定。当列为条目的几个事物有相同的名称时,最好加上括注,以利于区分。通用性原则一方面是指条名应使用规范或约定俗

成的名称,而不是杜撰的;另一方面是指当一事物有两个或两个以上名称时,应使用其中更加为广大读者知晓的名称。名词性原则是指标名必须为名词,理由在于条目作为注释的对象,主题是静止态的。即使有些事物原本为动态,但立作条目收入百科全书,整体上不是要做动态描述,而是说明科学内涵,它们也已经静态化。简要性原则是指作为注释对象的条目,转化为条头后是知识单元的“标识”,在形式上应当是凝练的,条名在表达上宜力求简要。上述条目化原则对于百科全书的选条十分重要。这些原则也成为衡量百科全书条目标名是否合理的标准。

(3) 条目是否能够通过多种途径检索。

百科全书的条目检索通过各种索引和参见系统来实现。索引量和参考条目数量是否充分、编排是否合理、质量是否有保证直接影响正文的利用率及百科全书的使用价值。

传统的内容主题索引分为简式和复式两种。简式主题分析索引是把索引条(包括条目标题和条目内容中隐含的主题)选出,并列地编排成表,用不同的字体或符号区别专门条目和隐含主题。复式主题分析索引是在按字顺编排的一级主题之下,把与其有关的主题作为二级主题集中编在一起,而这些二级主题又在其本身的字顺位置上作为一级主题单独出现,形成同一索引条目多处交叉并见的情况,复合分析索引就是带有交叉参见系统的索引。从普通读者检索使用上看,复式主题分析索引可使同一条目在索引中出现多次,为多途径检索提供方便,同时又能按主题集中相关资料体系中各主题的学科关系,弥补了按字顺排列而使得条目学科分散所带来的不足。但复式索引的编制却非常复杂,索引篇幅也会大大增加,是否采用复式索引要根据多个方面的因素来决定,如正文是否采用字顺索引、百科全书部头大小、其专科性如何等。因此,在判断百科全书索引编撰是否合理时,不能仅看它是否具有复合索引,有复式索引并不代表它的索引编撰是成功的,没有复式索引也不能片面地说其索引编撰失败,如百科全书正文若是按照知识分类来编排的,索引仍采用复式索引的形式就是一种浪费,是不恰当的。要根据百科全书本身的性质进行判断,根据索引本身的质量和数量进行判断。

百科全书参见系统繁简各异,有的提供双向参见,如《兰登书屋百科全书》,是最为完善的参见系统,充分揭示了内部条目间的学科关系。有的则从严精选,非必要时绝不参见,如《美国百科全书》(Encyclopedia Americana),这样避免了阅读本条目时注意力分散,效率下降。百科全书参见系统繁简度的评价是非常困难的,因为繁和简是一对矛盾体,过繁过简都是不合适的,而若能通过别的方式当简则简的同时又能化简为繁是最理想不过的了,第15版《新不列颠百科全书》是一个很好的范例,下文会集中阐述。此外,索引中的参见与正文中的参见有重复交叉,由于索引参见的范围更加广泛全面,正文的参见也应通过索引以便获得更加广泛的参考资料。

3)《中国大百科全书》信息组织的评价

《中国大百科全书》是中国第一部大型综合性百科全书,也是世界上规模较大的百科全书之一。第一版于1993年完全出版,全书按学科或领域分成74卷,包含66个学科知识领域,共收7.8万个条目,计1.26亿字,并附有近5万幅图片,内容丰富,1994年获第一届国家图书奖荣誉奖。《中国大百科全书》的第二版已于2009年8月由中国大百科全书出版社出版。

从现有文献对该书第二版的介绍可以看出,《中国大百科全书》第二版与第一版相比有以下三个方面的差异:①篇幅总体减少,卷数减为32卷(正文30卷、索引和附录2卷),6000万字左右,共计6万个条目,插图30 000幅,地图约1000幅;②重在普及。读者起点设定为高中文化程度,因此语言比第一版更通俗;③重检索。首先与第一版的分学科卷出版不同,第二版采用以汉语拼音为序、从A~Z的全字顺排列法编排条目;其次采用大、中、小条目相结合,以中、小条目为主的条目设置原则,中小条目占到了70%,最后,60 000个条目中几乎每个条目中至少有1个参见或索引,一些概述条目更是会涉及几十个参见或索

引条目,这将极大方便读者的查检。针对信息时代的特点,《中国大百科全书》第二版光盘版、多媒体数据版和网络版也已经出版发行。

从目前可以看到全文的《中国大百科全书》第一版来看,总的来说其编排体例是比较科学的。在条目编排方式上,编者没有照搬英美等国一概采用字顺法编排条目的传统做法,而是根据中国国情降低读者购买成本,将百科全书先按知识分类分卷出版,在具体卷册中各条目统一按条目标题的汉语拼音字母顺序并辅以汉字笔画、起笔笔形顺序排列。同音时则按汉字笔画由少到多的顺序排列,笔画相同的再按起笔笔形的顺序排列。第一字相同时,按第二字排列,以下类推。条目标题以拉丁字母开头的,排在汉语拼音相应字母部的开头部分,条目标题以希腊字母开头的,按希腊字母的习惯发音,分别排在汉语拼音字母的相应位置。另外,各学科卷在正文条目前一般有一篇介绍该学科卷内容的概括性文章,并附有反映该学科体系的条目分类目录。

在条目的选择及层级分工上,《中国大百科全书》采用的是大、中、小条目相结合的金字塔形的条目结构、以中小条目为主的条目设置原则,第二版的中小条目数量更是增加了不少。这种模式是立足于检索,兼顾阅读,符合现代读者的特点。对于较长条目,则在释文前列出了释文内的标题。对于标题层次较多的条目,则在释文前列出本条释文内标题的目录。对于那些学科间相互交叉的知识主题,编者则基本上做到各有关学科中均设置条目。例如,“马克思”在《哲学》卷和《经济学》卷及《政治学》卷均设有条目,但释文内容分别按其所在各学科的要求而有所侧重,使得同一主题的内容分散开来,这就要求读者在检索某一专有名词的时候最好先借助索引查找,以便检索到该主题的更广泛的信息。由于《中国大百科全书》是按学科分卷出版的,若读者手头只有某一个学科卷目,即使使用索引也无法获得该主题分散在其他学科卷里的信息,这也是该书信息组织的一个问题所在。

在检索系统方面,《中国大百科全书》做得完备而有特色,第二版更加强调了检索性。现有第一版的检索系统分为9个部分,读者可通过这9个途径进行检索。

(1) 音序检索。

各卷的条目均按汉语拼音字母顺序排列,采用汉语拼音音序法,为读者提供了最直接的检索入口。

(2) 笔画检索。

各卷均有“条目汉字笔画索引”,供不熟悉汉语拼音或不熟悉个别汉字读音的读者使用。

(3) 分类检索。

各卷正文前均列有“条目分类目录”,供读者从学科分类的角度检索条目之用,同时读者还可以从中了解该学科的知识构架,找到自己想要阅读的条目及相关条目。

(4) 内容检索。

各卷末均附有“内容检索”,这种检索除列有全部条目外,还列有条目释文中隐含的知识主题,所有内容索引的主题词数量相当于条目总数的4~7倍。一般来说,内容索引主题词数量在条目总数的3~10倍为合理范围。另外,各卷各条目释文中所出现的各种人名均附有生卒年,外国人名还全部附有原文,构成了该书最详尽的综合检索渠道。

(5) 外文检索。

除了纯中国内容的学科卷《戏曲·曲艺》外,其他各卷有附有“条目外文索引”,供熟悉和需要查阅外语的读者使用。例如,《天文学》卷总字数为1 540 000,共有条目1074条,除了纯中国内容而外文又无定译的32个条目未建立外文索引外,其他1042个条目均编有外文索引,占该卷条目总数的97%。

(6) 时序检索。

各卷大都列有各学科的“大事年表”,年表中所提到的人、事、物,凡有条目的均印

成楷体字，读者据此可比较方便地检索各有关条目。

（7）图片检索。

各卷均附有“彩图插页目录”，作为检索彩色插图之用。该索引虽然集中了图片，对读者全面检索有利，但彩图对正文的辅助作用却减弱了。

（8）文中参见检索。

当一个条目的内容涉及其他条目并需要由其他条目释文来补充时，采用参见的方式。各卷的参见系统由楷体字排印的“参见词”构成，在本条释文中出现的，则另用括号加“见”标出。例如，“秦二世胡亥即位后，对人民的剥削和压迫变本加厉，社会矛盾激化，终于在二世三年（前 209 年）激起陈胜、吴广领导的农民大起义（见陈胜、吴广起义）。”参见词把内容有关联的不同条目连接起来，相互贯通，从一个条目释文中可以得知本书所收的各种相关条目。

（9）书目检索。

各卷重要的条目均在释文后列出了有关的“参考书目”，向需要进一步研究的读者提供了详细的文献线索。

4)《不列颠百科全书》信息组织的评价

《不列颠百科全书》(Encyclopedia Britannica, 简称 EB, 又称《大英百科全书》)被认为是当今世界上最知名也是最权威的百科全书，是世界三大百科全书(《美国百科全书》、《不列颠百科全书》、《科利尔百科全书》)之一。《不列颠百科全书》第一版于 1771 年完成，共三册。经过多年的努力，在 1901 年美国出版商 Encyclopedia Britannica Inc. 买下 EB 的版权后，出版与编辑工作逐步转移到美国，现在人们熟知的大英百科全书公司已是总部位于芝加哥的美国公司。1929 年，随着第 14 版的问世，大英百科更投入了大量人力与物力，邀集近 140 个国家和地区的 4000 位学者专家参与撰述，大量收录欧洲以外地区的资料，完成全部 24 册的第 14 版，确立了它在百科全书界最崇高、最具权威的地位。目前印刷版中编排体例上最新的版本为 1974 年首版的第 15 版，共 30 卷，此后在 1985 年和 2007 年的更新版都对第 15 版的具体内容进行了更新。2012 年 3 月，EB 所在公司宣布不再发行纸质印刷版百科全书，该书将被彻底数字化。这一消息在国际出版界引起强烈反响。

从编排方式上看，EB 第 15 版不像过去那样全篇按照字母排列，而是分为《简明百科》(简称《简编》)、《详编百科》(简称《详编》)和《类目百科》(简称《类目》)三部分。《简编》采用中小条目编撰；《详编》采用大条目编撰，维持原来的大条目传统；《类目》取代传统知识分类的索引卷而肩负教育工具和分类索引两大职能。这种编排方式是一种全新的做法，极具特色，既兼顾了百科全书的教育与工具书功能，又弥补了传统按字顺分类带来的学科体系分散的不足。在具体每一卷中，采用的是字顺排列的方法，条目按照逐字母的方法排列。

在条目选择上，可以看出在 EB 第 15 版中，突出检索功能的《简编》在条目化方面要严于突出阅读功能的《详编》。在条目层级分工方面，EB 第 15 版的编者充分考虑到读者检索的 3 方面需求，即读者想查明的知识范围相当有限、读者查阅的知识范围较大但仍属于有限的课题(如想了解法国文学的全貌，而不是某部作品的出版年月或背景)、读者想从文化素养的角度获得对世界究竟意味着什么的了解，编者认为 EB 第 15 版必须针对这 3 种不同的检索要求，同时具有 3 种功能。因此，编者按照严格的条目层级分工，将全书分成《简编》、《详编》和《类目》3 个部分。其中用《类目》去满足第 3 种要求，同条目层级分工无关。但以小条目的《简编》和大条目的《详编》分别满足前两种检索要求，正是在实行条目的层级分工。《不列颠百科全书》曾经是大条目主义的典型，然而第 15 版改为通过《详编》和《简编》互补以实现层级分工，这意味着对大条目主义的否定。在国外的百科全书编纂中，大条目主义和小条目主义争执了近 200 年，目前则出现了两者互相靠拢以寻求补充的趋向。这无疑有

助于人们认识这两种主张的局限性和条目层级分工的合理性。《不列颠百科全书》第15版所采取的是大条目的《详编》和小条目的《简编》相结合的两极化的分工形式,这种编撰方式比较费时费力,《详编》加《简编》,一部书几乎相当于两部书。而《详编》和《简编》结合使用意在使阅读功能和检索功能并重,但实际应用过程中却并不适应当前读者的实际需求。因此,这种条目层级分工方法在理念上非常先进,但实际作用却有限。

在索引方面,EB第15版将索引变成小百科辞典,款目逐条注出其内容的简明解释,独立成书,再加上卷次和页码,使之成为其百科全书正文《详编》的索引,这就是EB第15版的《简编》部分。与传统索引形式相对照,EB第15版把索引编成百科辞典的形式,在一定程度上纠正了大条目主义不便查检和小条目主义不利于系统研究的缺点。

EB第15版的检索系统也比较完备,参见系统尤为有特色,检索途径有以下6种。

(1) 字母顺序检索。

各卷的条目均按字母顺序采用逐字母的方式排列,为读者提供了最直接的检索入口。

(2) 分类检索。

作为EB第15版分类检索途径,《类目》是一个由25 000个小条目组成的详细的分类索引,每一个条目都有3种对《详编》中不同范围材料的指引,既标出需要阅读的条目标题和所在卷页,还指出文中有关段落和其他有助于进一步了解的参考书目的卷页。

(3) 内容检索。

每条大条目前都列有“目次”,可帮助读者寻找他们所要研究的特定题目的具体内容,但EB第15版仅有按层次排列的章节标题,却没有标题所在页码,略有不足。

(4) 地图检索。

EB第15版中除了对正文条目做指引的索引外,还有专门的地图索引,地图索引按字顺把所有地图上出现的地名组织起来,对于在不同比例或不同区域的地图中重复出现的地名,索引都予以反映。它们或集中或分散,随地图的排列情况而定,帮助读者迅速找到地图中的某一具体地理位置。EB第15版采用地图分散随附索引的方法,使地图与正文紧密配合,有效地说明及补充正文,但不便于读者从特殊角度出发的寻检。

(5) 文中参见检索。

EB第14版和EB第15版的参见系统有较大差别。EB第14版的参见系统除了正文与索引的参见外,还有一套书目的参见系统,其参见范围最广,既有内部条目学科关系的揭示,又有对外界专著的引见和关联,构成了一个完整、丰富、开放的参见系统,便于读者对所查阅的主题获得尽量完整的知识。而EB第15版只采用索引参见,它以参见从严精选、非必要时绝对不参见为原则,甚至不在正文中做参见,虽然读者在阅读条目时不能随时了解有何相关资料,但却避免了阅读本条目时注意力的分散,如当确实需要参考时,也可通过《简编》提供详尽的参考途径。《简编》中包括许多指向《详编》的参见注释,《详编》提供大条目的详细信息,它们往往分许多章节,这些章节本身就足以在其他百科全书中构成独立的文章了,这种做法旨在使知识系统化,阐明各个内容之间的关系。

(6) 书目检索。

EB第15版参考书目按照条末集成的方法将参考书目附在条目最后,与正文密切相连,但其缺点在于使同一学科的书目被分散在不同主题条目之下。条目按作者字顺排序并标名著者、书名和出版年。

9.1.2 多媒体信息组织评价

多媒体技术是利用计算机技术将文字、图形、图像、声音等多种媒体综合一体化,使之

建立起逻辑连接，并能对它们进行获取、编辑、加工处理、存储和展示的技术。在非网络环境下，多媒体信息资源最主要的载体是光盘。本节选取光盘数据库为例，对非网络环境下多媒体信息组织进行评价。

1. 光盘数据库信息组织的特点

随着计算机存储技术的发展，光盘作为信息资源的载体具有存储能力强、介质成本低、数据可靠性高、便于携带、检索方便等优点，在 20 世纪 90 年代和 21 世纪初迅速普及，图书馆也将光盘数据库作为重要的馆藏予以采集、揭示和利用。按生产方式划分，光盘数据库可分为两种：一种是专业数据库商经过一次或二次加工整合形成的综合学术性光盘数据库，最初的学术数据库都是以光盘的形式存在的，如美国科技信息研究所（Institute for Scientific Information，简称 ISI）的三大引文索引数据库、美国《化学文摘》数据库、中国科学引文数据库、人大复印资料全文数据库等；另一种是将大型图书数字化并以光盘形式存储的数据库，如《四库全书》、《四部丛刊》、《中国大百科全书》、《不列颠百科全书》光盘数据库等。

光盘数据库虽然有很多优点，但也有其局限性。首先就是时效性不强，在信息检索中应用的光盘基本上都是只读式光盘（CD-ROM），数据库信息的更新依赖于出版厂商，现有的光盘数据库大多采用季度、年度等定期更新的方式，从而导致更新周期相对较长，不能满足时效性要求较高的检索的需要，如对学术信息的检索。其次是灵活性有所欠缺，由于光盘数据库主要采用的是菜单驱动的方式，使指令的运用得以简单化，因此，用户只要根据需要做出相应的选择即可，对于体例比较固定的工具书来说，浏览检索还比较方便，但对于经常需要更新系统和分类体系的综合性学术光盘数据库来说，这样的指令菜单使用起来灵活性不够。再次，光盘检索的成本不低，光盘信息检索价格低廉只是针对最终用户而言的。对于订购光盘的图书情报单位而言，光盘信息检索系统的价格仍然是十分高昂的。倘若一种光盘信息检索系统的利用率过低而更新速度又快，再加上需要投入较昂贵的辅助设备，其成本也是很高的。最后，虽然与传统参考工具书相比，光盘信息量很大，但仍然是有限的，再加上光盘介质不易长期保存，这对图书馆来说也是一个巨大的负担。

因此，近些年来，光盘数据库特别是学术性光盘数据库渐渐被网络版数据库所取代（如 ISI 的三大引文索引数据库光盘被其基于网络的学术资源整合平台 Web of Knowledge 取代），但大型图书的数字化光盘数据库却没有退出市场，依然占据着重要的地位。这是因为这种类型的图书涵盖的信息量大、知识价值高，但由于部头大、携带不便，手工检索的效率不高，采用光盘数据库的形式存储和运用各种信息技术对这些图书数字化加工可实现多途径全文检索，提高读者的检索效率。同时，这类图书的更新周期比较长、体例设置相对稳定，也不会因为光盘数据库的时效性不强和灵活性欠佳而有所影响。另外，这类图书的纸质版购买成本往往大大高于光盘版，如《中国大百科全书》（第二版）的纸质版价格为 8000 元左右，而中国大百科全书出版社出版的《中国大百科全书》光盘 1.2 版 DVD 售价仅为 98 元，相差 80 倍左右。不仅如此，图书馆还可通过光盘联机检索提高其利用率，最大限度地节约购买成本。因此，目前经常使用的光盘数据库主要是图书数字化的全文数据库，下文讨论的光盘数据库信息组织的评价也将以此为对象。

2. 光盘数据库信息组织的评价指标

光盘数据库的信息组织的优劣与否主要通过其检索功能来体现，要衡量光盘数据库的检索性能是否完备、检索手段是否快捷、检索入口是否实用，就需要有一套可广泛适用的评估指标，以此判断光盘数据库在检索方式上的优劣。光盘数据库信息组织的评价指标主要有以下 4 个方面。

1) 检索途径

所谓检索途径是指选择哪些字段或数据元素为检索点,是否能持续而方便地对各检索点进行检索,它是查看光盘检索系统提供的检索入口。检索途径多是光盘版优于印刷版的一个显著特点,其根本目的就是为了满足人们利用多种途径进行文献检索获取原文信息,提高检索效率。目前,检索途径包括主题、分类、著者、题名、号码、年代等浏览检索途径,较为先进的还提供全文检索功能。总体而言,检索途径越多越利于快速而准确地进行检索目标定位。工具书光盘数据库更应利用自身优势,将原文中的各种目录、索引、参见系统等挖掘出来,运用超链接的技术进行互连,以方便读者使用。

2) 检索策略

光盘数据库的检索策略包括两个方面。一是对检索符的支持,是指光盘数据库的检索是否支持布尔逻辑检索、邻近检索、截词检索、字段限制检索、短语检索、括号检索、自然语言检索、多语种检索、模糊检索及区分大小写检索等。对用户而言,检索符多样并可以以复选框的形式进行选择是最方便的。具体来说,检索文本框下拉列表中是否有“布尔逻辑式”等选项以方便用户随时选用,“AND”(“与”检索符)是否自动添加到用户输入的检索词之间,等等。

二是检索模式的应用,是指光盘数据库是否运用了多种检索模式;设置简单、高级等何种检索方式为系统默认值;是否支持族性检索、能否运用组配操作进行更专业的限定检索;能否在下拉列表框中设置相应专业领域的检索入口,以将用户检索限定在特定领域中;是否支持自然语言的提问方式的检索;等等

不同光盘系统中这些检索策略的应用形式是不同的,如有些系统要求用提问表达式输入;有些系统则用菜单方式,显示一个工作单,工作单上的各行设计成各个字段的专用输入行;有的布尔逻辑运算符用简单的命令输入,有的则用隐含的方式提供。无论如何处理,用户对检索策略的要求通常是输入要求清楚、命令简单明了、按键次数少。有的系统还提供保留检索策略和修改检索策略重新执行检索的功能。

3) 操作方法与操作界面

操作方法有指令式和菜单式之别。指令式的操作方法中,用户根据检索需求及检索规则输入相应的检索式,指令式操作从简单到复杂,差异很大,它对操作者检索专业素养往往要求较高,如世界上最大的联机检索系统 Dialog 复杂的检索方法及高昂的联机查询费用使得其检索长期以来一直依赖于专职情报工作人员的操作。而菜单式检索直观性强,易操作,采用鼠标选择,减少了键盘输入的工作量,用户一般也不需要去记操作指令,光盘数据库采用该方法将大大提高用户的检索效率。现在主流的光盘数据库采用的均是菜单式的操作方法。

除了检索性能外,操作界面也是衡量光盘数据库优劣的因素之一。操作界面是指通过菜单、图标、模块、功能键或命令等操作完成人机交互,它是用户进入光盘检索系统后第一眼就能看到的東西,它给出一些基本信息,如当前驱动器中光盘数据库收录时间、范围、版权、数据库内容说明、系统情况及命令。一套较优秀的光盘库,应该设置原版全文浏览器,内含原版浏览模块和原版打印模块,具有自动编辑剪贴功能,可通过相应的检索软件检索和浏览全文信息(包括文本、图表、照片等),其版面显示既可与原刊一致,也可以按照整套系统的版式特征统一规划显示文档内容。考虑较周全的系统还在屏幕下端给出用大写字母印出的当前功能,使用户知道系统正准备做什么。总的来说,友好的操作界面应是直观性强、便于操作并提供人性化的帮助界面。

光盘的操作界面大多是友好的菜单式结构,有的菜单不仅可以作为其他菜单的子集出现,还可以从特定的命令级中产生。有的屏幕上可以同时显示多个窗口,而新的窗口又可以相应用户的命而现。一般在当前菜单或窗口中可以显示调用其他相关窗口或菜单的命

令, 于是可以随心所欲地使用各种功能。

4) 辅助功能

光盘数据库的辅助功能的设置应以最大化地方便用户为目标。一些常见的辅助功能包括检索记录的保存、检索结果的显示、打印和转存等。数据库是否提供了相关的便于检索和处理检索结果的功能? 这些功能是否完善? 如可否按用户的要求将检索结果显示在屏幕上、直接打印出来或复制到存储器上? 系统是否允许用户提出输出的格式和排序的要求? 用户是否能够通过检索历史记录保存来浏览上次检索的结果并在新一轮的检索中对以往的检索方式和策略进行相应的调整?

辅助功能中还应有良好的帮助功能, 帮助内容是否容易调用、是否详尽丰富、是否有可查检的帮助内容索引, 是衡量帮助功能是否完善的标准。例如, 在操作过程中只要按“F1”功能键, 就会立即看到有针对性的“帮助”屏幕, 使初次使用者也会感到很方便。

3. 光盘数据库信息组织的实例

百科全书的分类结构及用户对其检索性的高要求使它非常适合采用数字化的光盘这一载体。最近几年, 大多数主要的百科全书都已经在不同程度上电子化, 基于 CD-ROM 的百科全书拥有携带方便、成本低廉的优点。同时, 电子百科全书还可以包含各种传统媒体无法承载的多媒体格式, 如动画、音像或视像。更重要的是, 光盘版的百科全书可以实现相关概念间的相互动态链接和用户检索的最大便利。

1) 《中国大百科全书(简明版)》光盘版简介

《中国大百科全书(简明版)》是在《中国大百科全书》第二版出版之前, 为满足更多读者的需求, 由中国大百科全书出版社组织编纂出版的。它的知识体系以《中国大百科全书》为基础, 包括历史、地理等 75 个学科或知识领域。它删除了《中国大百科全书》各卷中重复、交叉、过专、过僻的条目, 增补了《中国大百科全书》欠缺的知识总论、国家、能源、材料、信息、旅游、民俗及服饰、烹饪、家政等方面的条目, 更新了《中国大百科全书》的资料和数据。它充分反映了近现代尤其是近 20 年来中国出现的新科技、新事物、新情况和新成就。

《中国大百科全书(简明版)》光盘版是以《中国大百科全书(简明版)》为文本并加上多重检索功能编辑而成的。光盘共 2080 万字, 包括条目正文、附录和索引三大部分。条目正文包含 31 000 个条目及相应的插图或表格 11 000 幅, 内容注重深入浅出, 力求简明易懂; 注重介绍具有普遍性、典型性、实用性和高检索率的知识; 注重引用资料 and 数据的时效性、准确性和权威性; 随文所附的插图和表格编辑严谨、直观生动。附录为“中外大事年表”, 高度概况地介绍了人类文明发祥以来全人类各时期、各地区、各领域发生的重大事件, 揭示了人类文明的发展脉络, 为光盘使用者提供了社会发展的基本线索。索引包括对条目标题的全文检索、音序检索、笔画检索, 以及文内热字跳转查阅检索等, 内容即检即现。下面以《中国大百科全书(简明版)》光盘版为例来评价光盘数据库信息组织的特点。

2) 《中国大百科全书(简明版)》光盘版信息组织评价

根据上文所提出的光盘数据库信息组织评价标准, 《中国大百科全书(简明版)》光盘版的信息组织可从以下 4 个方面进行评价。

(1) 检索途径方面。

《中国大百科全书(简明版)》光盘采用超文本链接技术来体现百科全书承载的知识体系结构, 采用压缩技术把庞大的数据压缩到一张光盘上, 在此基础上提供方便快捷的条目检索和条目间的参见跳转。

它所提供的检索途径有条头(条目标题)的全文检索、音序检索、笔画检索及文内热字

跳转检索。条目标题全文检索功能可分别在音序检索界面和笔画检索界面实现,在这两个界面中均设有条目标题的检索文本框,如图 9.1 和图 9.2 所示。严格来说,该光盘数据库的音序检索界面并没有实现音序检索功能,只是将条目按照音序的方式进行排列,用户从中选择。真正可按音序检索的是光盘操作界面上的 23 个字母,如图 9.4 所示,单击某个字母就会出现以该字母开头的所有条目中的首条条目,通过向后跳转可查看以该字母开头的其他条目,但并不提供该字母开头的所有条目的整体索引,因此使用也有所不便。笔画检索提供按首字笔画检索浏览的功能。如图 9.3 所示,文内热字跳转检索是指将正文条目及“中外大事年表”中涉及与条目正文相关的知识主题参见用绿色字体标识出来,并对之进行超链接,用户单击即可跳转到热字对应条目,可以连续跳转,也可顺序返回或直接返回。只有正文条目可用以上四种方式检索,“大事年表”和“数字与字母”没有实现前三种检索功能,用户只能按时间顺序逐个浏览。

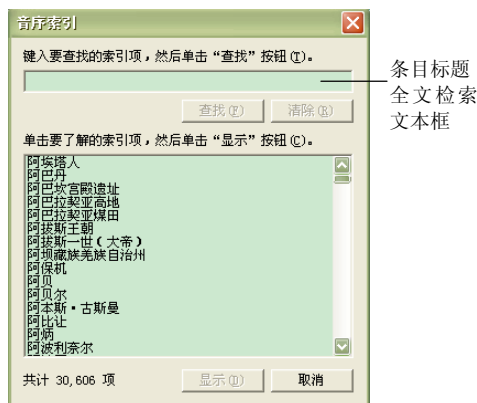


图 9.1 《中国大百科全书(简明版)》

光盘版音序检索界面

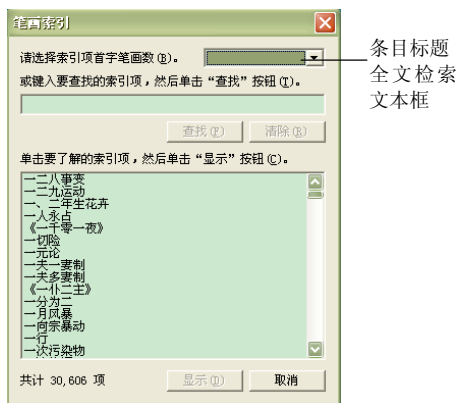


图 9.2 《中国大百科全书(简明版)》

光盘版笔画检索界面

(2) 检索策略方面。

该光盘数据库不支持任何检索符,也没有多种检索模式可供选择,所有的检索都采用首字匹配的方式检索。由于综合性学术数据库中资源著录字段繁多、资源类型多样,考虑到用户的多样化需求,它们比较注重检索模式的多样化设计和检索符的组配支持。而该百科全书光盘数据库的内容条目构架比较简单,由条目名称和释文两部分组成,在使用上文所述的四种检索途径时若采用繁复的检索策略还会增加用户使用上的困扰。但该光盘数据库的全文检索仅局限在条目标题范围内,而没有普及到整个条目全文系统,不能不说是一个很大的遗憾。若整个光盘内容都可进行全文检索,多种检索模式的设计及检索符的综合运用则是必要的。

(3) 操作方法与操作界面。

该光盘数据库的操作界面比较简单易用,共分为 7 个模块,如图 9.4 所示。

进入该界面前还有一段将我国古代四大发明、古今中外著名人物事件交织起来的 Flash 宣传片,以彰显该书内容的广博精深,引人入胜。图 9.4 中的“出版信息”模块包括出版社简介、编委会介绍、《中国大百科全书》(简明版)光盘版及纸质版的前言和凡例、学科顾问和主要编创人员等基本信息。

在操作方法与操作界面上,该操作系统采用菜单式操作方法,界面友好,设计简洁明了,用户可直接通过这个界面进行检索和浏览,而不需要再层层搜索。在浏览具体条目、“中外大事年表”、“数字与字母”模块时,可以实现各模块功能的菜单式切换,如总目录、音序索

引等，如图 9.5 所示。但美中不足的是，屏幕上只能显示一个模块窗口，新的窗口打开前必须关闭前面打开的窗口。

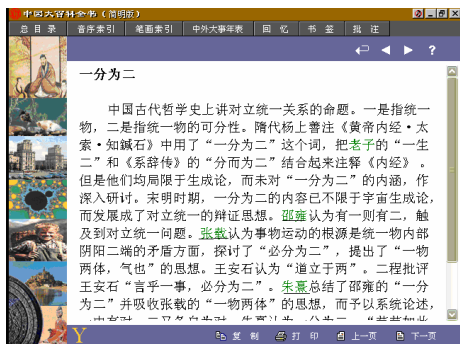


图 9.3 《中国大百科全书（简明版）》
光盘版热字跳转示例^①



图 9.4 《中国大百科全书》（简明版）
光盘版操作界面

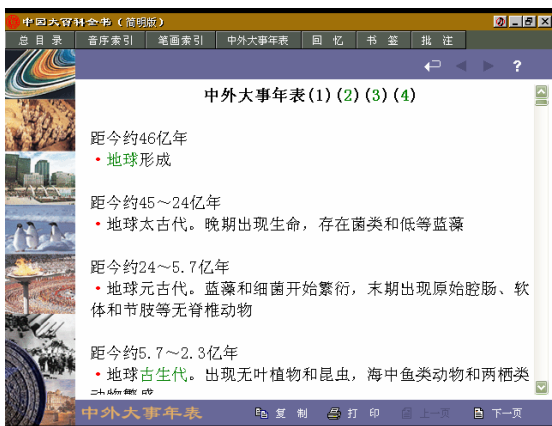


图 9.5 《中国大百科全书》（简明版）光盘版“中外大事年表”浏览界面

（4）辅助功能。

如图 9.5 所示，该光盘数据库设有打印、复制、回忆、书签、批注、返回和前后跳转这 7 种功能，还是比较丰富的。

单击“复制”按钮，当前条目内容即被复制到剪贴板，打开任何一个文字处理软件，粘贴即可。单击“打印”按钮，当前条目内容即被打印出来。打印和复制功能均是对整个条目正文页面而言的，用户不能随意选择其中的某一部分进行操作。回忆功能保存了用户本次运行光盘后的浏览和检索记录，用户利用回忆功能可得到其阅读期间近期所查看的所有条目名称及顺序，并可在其中选择某个条目，直接查看。该光盘设有书签功能，用户可以将需要标记的页面或词条定义自命名书签，单击书签标签名可以随时查看，不用再逐一查找，书签定义后被保存在用户硬盘中，直到用户删除该书签为止，使用非常方便。批注中包括粘贴、

^① 为表现热字的显示度，笔者将热字下都加上了下划线，光盘中热字采用绿色字体标识。

复制和删除功能,供用户在阅读时做批注,用户可把阅读某个条目后的感想及需要保存的其他信息加入批注内,批注也是保存在硬盘中的。但它没有提供检索功能,用户批注过后必须通过加注标签才能标识批注所在位置,方便下次查找。返回功能是指返回上一个浏览记录,前后跳转是指可以查看该词条的前后词条功能。

它的帮助系统共分三部分:第一部分介绍整个界面的总体情况和各模块的作用;第二部分是对具体如何使用音序索引的介绍;第三部分是对具体如何使用笔画索引的介绍。这个帮助系统做得非常简单,没有详细的帮助目录索引和文字叙述,只是用图示的方式标明各部分框架的内容和作用。用户可在使用过程中遇到问题时随时单击“?”图标以查看帮助。

另外,该光盘数据库中只包含了非常少的多媒体资源即背景音乐和片头动画,它并没有充分利用光盘介质可容纳多媒体资源的优势为用户提供形象、生动的阅读服务。

4. 光盘数据库与网络版数据库信息组织的比较与实例

光盘作为制作数据库的一项主要技术,曾给传统的文献检索带来巨大的冲击。光盘服务器和光盘网络技术的日趋成熟,使光盘数据库实现多种检索及多用户共享。但随着网络技术及数据存储技术的发展,越来越多的出版商开始把题录、文摘、全文光盘数据库等信息产品转移到网上发布。例如,国外的EBSCO全文数据库、ProQuest光盘库,国内清华大学的中国知网、中国科技信息研究所的万方数据资源系统等,网络版数据库渐渐成为学术数据库的主流。

1) 光盘数据库与网络版数据库信息组织的比较

尽管光盘数据库在多用户数据库领域的地位逐渐削弱,但不会消亡。网络版数据库由于网络媒体的特点,越来越成为目前数据库尤其是学术型数据库中的主导。光盘数据库和网络版数据库作为目前数据库的两种载体形式,各具特色。二者在信息组织方面的相同之处有三点:① 检索实质相同,它们都采用一定的检索手段,依据特定的检索标准对入库文献和用户的检索提问进行标引,并将二者进行比较以寻找数据库内与之相匹配的信息;② 逻辑组织大体相同,它们在逻辑布局上的核心部分都可分为文献信息库及其索引数据库两大部分;③ 逻辑组配手段相似,二者为用户提供构造检索表达式的逻辑方法都是一样的,支持短语检索、布尔逻辑检索等多种检索符。二者的不同之处主要体现在以下几个方面。

(1) 检索界面。

目前的光盘产品多没有统一的标准,各自的内容、结构及检索界面都不同,操作烦琐。早期的联机检索系统的检索方式、输出的结果都是特定的检索命令,检索人员一定要熟知该检索系统的命令格式,否则就会浪费检索时间和金钱。随着检索实施者由专门的检索人员逐渐转为一般的信息最终用户,网络版数据库的检索系统的检索界面越来越友好,检索功能和方法相对光盘要简单些。尽管各数据库界面不尽相同,但其基本检索方法大致相同,一般是主题词或分类的单层次检索,使用起来比较容易。

操作界面大多选用简单容易操作的菜单式检索方式,在菜单检索窗口列出功能键和功能词,即使从来没有接受过专门训练的信息用户,根据提示,一般都能完成检索任务。而且一般的网络版数据库都有优秀的在线检索帮助,根据在线帮助,有助于用户进一步了解数据库,学习更好的检索方法,如检索字段和正确的输入格式等。

(2) 检索功能。

尽管网络版数据库在其发展初期与当时较为成熟的光盘数据库相比,检索方法过于简单,如一般只有主题词检索和分类检索的单层次检索,也很少具有逻辑检索,即“与”(and)、“或”(or)、“非”(not)的组合检索和嵌套检索、模糊检索和智能检索等多层次的检索功能,其检索途径和检索记录所包含的字段也比较少。但随着网络版数据库的发展,光盘数据库,

特别是学术光盘数据库大多都被网络版数据库所取代,网络版数据库的检索功能克服了上述问题,达到了较高的水平,如大型的学术资源整合平台——ISI Web of Knowledge。

但光盘数据库也有其独特的检索优势。首先,由于光盘数据制作严谨,光盘检索比网络检索有更高的准确性,光盘数据库的索引制作也较为完整,用户可通过各种索引浏览和了解整个光盘数据库的内容。再者,光盘有较高的安全性,由于光盘具有只读性,因而比其他信息介质更具安全性,被病毒侵袭的可能性最小。

(3) 检索效率。

由于不同的光盘数据库的收录对象具有特定性、侧重性,用户为了某一个特定的研究课题往往要检索不同的数据库,既费时又费力。而像 DIALOG 等国际联机检索系统具有同时检索多个数据库的功能,但是联机检索系统的费用较高,检索策略较为复杂,在国内仅限于专业人员的检索使用。而网络版数据库则集多个检索系统于一体,如 ISI Web of Knowledge 就包括了 ISI Web of Science、Conference Proceedings Citation Index、INSPEC、BIOSIS Previews 等在学术界颇有影响的文摘型数据库。利用网络版数据库中的导航功能,能为用户提供该数据库与其他相关数据库资源的信息检索。同时网络版的数据库还可以与电子期刊全文链接,在检索的基础上直接在线浏览全文,从而避免了像光盘版数据库那样检索到的文摘还要再到其他的数据库检索原文的情况。这些整合和超链接系统都大大节约了用户的检索时间,提高了用户的检索效率。

(4) 数据库的通用性及使用要求。

光盘数据库的兼容性差、维护困难。它需要专业技术维护人员,服务器上也要定期更换光盘和更新检索软件,同时还要考虑系统是否支持不同的计算机平台和操作系统。网络版数据库的通用性较好,它传递信息的基本模式为:存储在中心服务器上的数据、搜索引擎、索引引用支持软件,用户在 Web 界面通过网络从服务器上访问数据。网络技术解决了许多在光盘检索中存在的问题,具有光盘无法比拟的优势。网上信息产品比光盘网络更容易支持,互联网的系统汇集性使得图书馆可以建立支持多种系统、提供多样服务的标准网络,如果该网络能支持 TCP/IP 网络传输协议,且在每台客户机上有 Web 浏览器,那么图书馆就具备了支持任何网络信息产品的条件,包括网络版数据库。

一般的光盘数据库检索到的文献信息要根据该数据库的要求保存到一定的文件夹或指定的服务器上,在特定的 IP 地址进入指定的地点才可以浏览到自己检索的文献信息,具有极大的局限性。而网络版的文摘索引型数据库则可以把检索到的文献保存在本地硬盘或软盘中,以后用户在任何一个联网的计算机上,都可以看到自己以前所检索到的文献资料,快捷、方便,不受时空的限制。

(5) 数据库存储容量和更新速度。

在光盘数据库中,一年的文献数据一般需要两三张光盘,用户要检索多年跨度的文献,需要多次更换光盘才能完成检索任务,而由于参考文献、引用文献和相关文献信息往往存放在其他光盘上,用户需要多次反复后才能看到完整的引用和被引用的文献和相关的文献信息。而网络版数据库由于网络本身不像光盘那样会受到自身容量的限制,其信息量随着网络范围的不断扩大、网络技术的发展、人类知识的提高而与日俱增,拥有更加庞大的数据存储容量。它可以将多年的数据集中于一台服务器,相关文献之间便可以实现无障碍地互相连接,用户只需要使用鼠标点击即可。

毋庸置疑,网络版数据库还有更加高效的更新速度。作为信息传递介质,网络的时效性大大优于光盘,网上的信息是动态的,网络版数据库数据每时每刻都可更新。而光盘数据的时滞较长,一般在半年以上甚至更长,大大滞后于用户检索需求。

2) 网络数据库实例分析

Web of Knowledge (以下简称 WOK) 是由 Thomson Reuters 公司推出的基于互联网的新一代学术信息资源整合体系。它以 SCI、SSCI 和 A&HCI 三大引文索引数据库为核心整合多种学术资源,并结合信息资源分析工具和信息管理工具为用户提供学术信息服务,是目前运用最广泛的权威网络学术数据库,其数据库内容所覆盖的学科领域也非常广泛。多种资源整合及一站式检索是该平台信息组织的两大特点。

(1) 多种资源整合。

WOK 是一个基于 Web 而构建和整合的数字研究平台,通过强大的检索技术和基于内容的连接能力,将高质量的信息资源、独特的信息分析工具和专业的信息管理软件无缝地整合在一起,兼具知识的检索、提取、分析、评价、管理与发表等多项功能,从而大大地扩展和加深了信息检索的广度与深度,加速科学发现与创新的进程。

在信息资源整合方面,WOK 以 Web of Science(整合三大引文索引数据库 Science Citation Index Expanded, Social Science Citation Index, Arts & Humanities Citation Index 而成)为核心,凭借独特的引文检索机制和强大的交叉检索功能,有效地整合了各类型的学术资源。它们分别是:学术期刊 (ISI Web of Science, Current Contents Connect)、学术专著 (Current Contents Connect)、会议录文献 (ISI Proceedings)、发明专利 (Derwent Innovations Index)、化学反应 (Current Chemical Reactions, Index Chemicus)、Internet 学术资源 (External Collections) 及其他多个重要的学科数据库 (BIOSIS Previews, INSPEC, FSTA, PsycINFO),提供了自然科学、工程技术、生物医学、社会科学、艺术与人文等多个领域中高质量、可信赖的学术信息。该体系突出的特点是以 Web of Science 为核心,不仅建立起了包括期刊、专利、会议录在内的多种类型文献之间的相互引证、相关参考的关系,而且还实现了对拥有使用权限的全文文献及事实数据 (如 GenBank, ISI Chemistry Server) 的链接。这种对资源的整合构成了一个动态的学术信息门户,可以全方位地为科学研究提供文献信息保障,使科研工作者得以了解与其研究领域相关的各种类型文献,以及学科过去、现在和将来的脉络发展与交叉。

WOK 还整合了专门的学术分析评价工具“期刊引证报告”(Journal Citation Reports, 简称 JCR)和“基本科学指标”(Essential Science Indicators, 简称 ESI),用以帮助用户评价所需信息资源的质量和影响力,帮助研究人员迅速深入地发现自己所需要的信息,把握研究发展的趋势与方向。此外,该平台还设有免费检索工具 ISI Highly Cited.com,供用户查找近年来在某个学科领域内做出重大贡献的学科核心人物,其内容包括这些人物的传记和出版物信息。目前,入选的有神经系统科学、工程学、物理、化学、计算机科学、地球科学、分子生物学、遗传学和空间科学等 21 个学科领域,计划每个学科挑选 250 名重要研究者入选。

Reference Manager、ProCite 和 EndNote 都是 WOK 平台上的信息管理软件,为用户查找、组织和格式化其检索的信息资源书目带来了极大的方便。它们不仅能够帮助用户管理由 ISIWOK 检索得到的文献信息,也可以管理由其他系统或平台检索到的文献资料,包括个人收集的参考文献,用户用该类软件可以自建一个完全个性化的参考文献数据库。

此外,WOK 还提供了著作管理与科研社交工具 ResearcherID,为全球研究者打造免费科研交流平台。2012 年 10 月,WOK 推出数据引文索引 (Data Citation Index™, 简称 DCI),DCI 作为 WOK 平台上一个新的研究资源,将推动对数据集和数据研究的发现、使用及归属,并把这些数据与同行评议文献连接起来。

(2) 一站式检索。

在多种资源整合的基础上,WOK 提供了强大的检索功能,包括跨平台的一站式检索及基于内容与引文的跨库交叉浏览等。

如图 9.6 所示,Web of Knowledge 的检索功能非常强大,它不仅可以进行单个数据库资

源检索,还可根据需要选择参与交叉检索的数据库实现跨库检索,同时辅以信息资源的检索、提取、管理、分析与评价等多项功能,从而大大扩展和加深了信息检索的广度和深度。

检索结果可按相关度、出版日期、第一作者名、被引频次、来源出版物、入库时间等多种标准排序,如图 9.7 所示。值得注意的是,WOK 检索结果页面的每条检索记录下都会有对应的图标表示该记录所在的数据库,用户通过单击图标进入不同的数据库浏览这条数据。这种方式虽然为用户提供了多种获取途径,但普通用户对各个数据库的收录特点并不熟悉,并不知道哪些数据库是收录全文的,哪些是收录文摘的,在反复的单击查看中降低了检索效率。现在使用“全文”(Full Text)图标,表示该记录提供全文链接,体现了该平台信息资源整合的趋向,同时这也更加符合用户一站式获取信息的需求。



图 9.6 Web of Knowledge 一站式检索界面



图 9.7 Web of Knowledge 检索结果的显示界面

用户选定检索结果后,可以通过各种超链接方便地获取有关该记录的信息,如图9.8所示。图9.8显示,该平台可提供的超链接主要有10种,分别介绍如下。

(1) 全文链接(Full Text)。在记录中,如果该文献在WOK中有全文,便会出现供浏览全文的链接,只要点击此链接,用户便可直接看到当前记录的全文,但前提是用户的图书馆同时购买了该论文所在期刊的电子版,由WOK替用户进行“全文”链接。

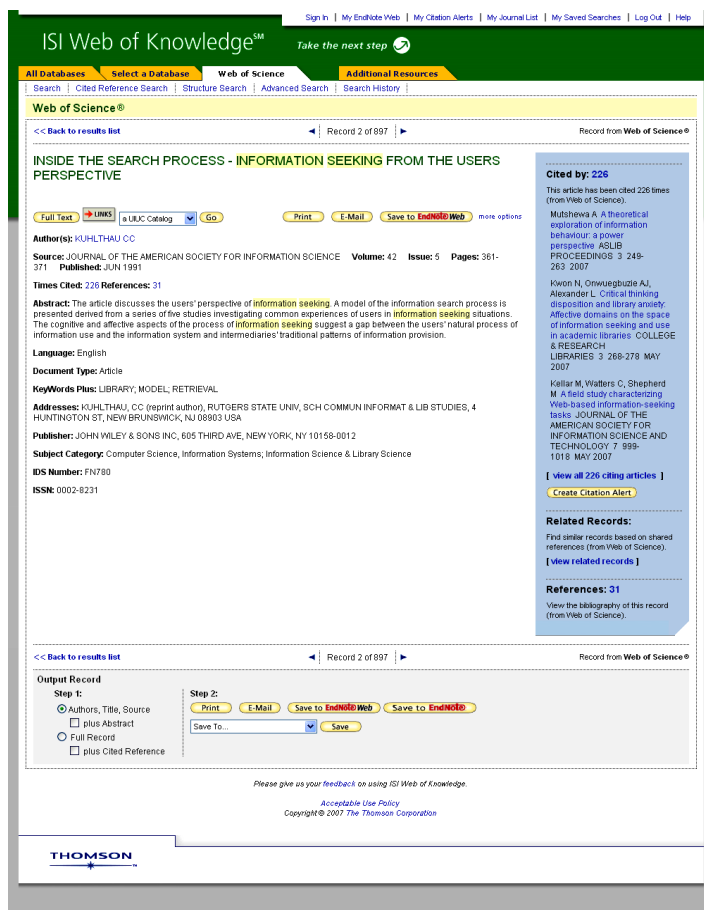


图9.8 Web of Knowledge 某一条检索记录的显示界面

(2) 引文链接(Times Cited)。它显示的内容包括当前记录的被引频次,还提供“创建引文跟踪”(Create Citation Alert)功能,方便用户追踪该论文的引文信息。

(3) 参考文献链接(Cited References)。该链接指向的是本条记录引用的参考文献,并提供“引证关系图”(Citation Map)。

(4) 相关记录(Related Records)。通过“相关记录”链接,用户可以查看在不同年份中与当前所检索的记录共同引用同一篇或几篇参考文献的一组论文,并按共同引用参考文献的多少(即相关度)排序,和当前记录引用的相同文献越多,该文献与当前记录在主题上越靠近,则该文献在列表中的位置就更靠前。而且,WOK对于“查找相关记录”的检索结果数不加任何限制,命中几条就显示几条。

(5) “Send to ...”模块。该模块可以将检索记录发送到 my.endnote.com、EndNote、ResearcherID 等文献管理工具中进行,有利于用户组织、管理文献资源并应用于论文写作。

(6) 与多种图书馆馆藏目录系统的连接。在全文链接下方的下拉列表框中选择本馆的馆藏目录即可连接到本地的 OPAC 系统, 找到该文献所在期刊的馆藏记录。

(7) 与 JCR Web 的链接。研究人员可以通过文献的被引率 (Citation Rate) 和期刊的影响因子 (Impact Factor) 迅速了解科学研究的相对影响, 为科研绩效的评价提供了科学的量化依据。

(8) 与 ISI Highly Cited.com 的链接。有助于研究人员迅速了解本领域有影响的专家及其成果与联系方式。

(9) 与 Essential Science Indicators 的链接。该库可以用来分析各个领域学术研究的发展、影响和趋势。用户可以从该数据库了解达到一定级别的科学家、研究机构 (大学)、国家 (城市) 和学术期刊在某一学科领域的发展和影响。

(10) 与网络资源的链接。通过站外资源 (Additional Resources) 的链接, 可帮助研究人员找到与课题相关的网站述评 (Web Site Reviews)、研究活动 (Research Activities)、预印本 (Preprints) 和资助项目 (Funding)。

过去使用的各种数据库都以一种零散的、孤立的状态存在着, 即使若干个库捆绑在一起, 也仅仅局限在使用同一界面层次上, 体现不出文献内在的相互联系。WOK 则建立了不同类型资源之间的关系, 最大限度地保持知识体系的完整超链接的广泛运用, 使之成为一个有机的整体, 从而消除了由于数据库收录范围有限而造成的知识体系的割裂。这种对资源的整合可以全方位地为科学研究提供文献信息保障, 使科研工作者得以了解与其研究领域相关的各种类型文献。

9.2 网络环境下的信息组织评价

9.2.1 网站信息资源组织评价

网站是网络信息资源的重要组成部分, 是一种用标记语言 (描述性语言) 将信息组织好, 再经过相应的解释器或浏览器翻译出的包括文字、图像、声音、动画等多种信息的组织方式。在目前的网络环境下, 网站建设中一般都会糅合、混搭 Web2.0 的相关应用与技术, 如 RSS、Tag、SNS 等。关于 Web2.0 环境下的信息资源组织评价问题, 我们将在 9.2.5 节中专门论述。

1. 网站信息资源组织的特点

随着计算机和网络的普及, 网络信息数量和种类都以指数形式增长, 网站特别是门户网站的用户数量也在不断增加。为了对这些急速增长的网络信息资源进行有效的组织以满足不断增长的网络用户需求, 网站信息资源组织呈现出以下一些特点:

1) 组织对象的数量庞大

从网络的组织结构可以看出, 信息资源主要分布在网站上, 网站作为网络信息与网络用户之间的中介, 其最终目的在于将网络信息有序化、整合化, 向用户提供优质服务。网站尤其是综合性门户网站的用户群体广泛, 面对互联网上海量的信息必须分设栏目, 争取做到“面面俱到”。因此, 网站信息与其他类型的网络信息组织形式相比, 它所组织的对象是最广泛的, 这给信息的组织和保存带来很大的困难。

2) 信息质量不易控制

网站由一个主页 (Home Page) 和若干个从页 (Web Pages) 组成, 每个页面都充斥着大量的信息, 这些信息还要随时更新。而网站建设的人力和资本有限, 因此, 面对大量的信息其对信息资源质量的控制也是非常有限的。加上互联网的开放性, 使得大量虚假信息、淫秽

信息、色情信息、暴力信息和进行煽动性政治宣传的网站充斥着互联网,这更使得网站信息质量的控制难上加难。

3) 信息资源组织体系不易规范

网站用户群体的广泛性和网站经营的商业性给网站信息资源组织标准化带来困难。正因为用户知识层次和需求的多样性,网站信息资源的分类系统不可能按照标准的学科分类法系统来进行组织,而网站的商业化运营也使得网站不可能为每条信息进行严格的质量控制和规范化著录。

2. 网站信息资源组织的评价标准

网站的评价标准体系有多种。根据互联网网站建立、构成要素、信息传递与服务的独特性,网站在组织信息中的评价标准体系及其具体指标应该包括以下几个方面。

1) 主页的设计与安排

主页是在访问某一网站时访问者所接触到的第一个网页,这一网页起着引导作用,引导访问者访问与其链接的相关网页。对任何网站而言,主页是最重要的网页,因为它是整个网站内容的“指南针”。主页的内容组织与结构安排对网站的性能和网站信息的利用至关重要。

要合理地规划与组织主页的内容,首先要确定网站设计的目标与用户定位。要考虑使用者特性,如背景、习惯与能力。互联网的发展将从“吸引眼球”向重视服务转化,信息的质量、易用性及其广度和深度等已逐渐成为用户选择、利用网络信息的重要标准。这些都对网站的设计与管理提出了更新、更高的要求。在对网站主题进行规划时,切忌使网站的主题过于分散。因为根据搜索引擎的判断,网站主题越集中,一般情况下网站所有者在这方面投入的精力会越多,因此所提供信息的质量也会越高。

主页的设计是否一目了然?主页上的链接项目是否只限于几个高级的类别(6~8个最为理想)?颜色数量是否太多?图像是否过大?画面是否过于复杂?是否充分利用多媒体技术?网页的文字、图片、动画、音乐、程序等成分的处理是否妥当?(如空间安排、文字大小、标题的醒目、重要内容的突出,颜色、图形与符号的使用都要避免太过或不够。)整体的网页风格是否与主页具有一致性?

网站主页是否包含几个必备事项?必须包括标题、站点介绍(About Us,关于我们)、提供的产品或服务、主要栏目或内容、网站地图(Site Map)或网站概览、版权资料、电子邮件地址或其他联系资料等。

2) 组织结构

组织结构是网站信息组织的框架与呈现方式,反映了对网络信息进行整序、优化,并集成为一个便于有效利用的系统的方法。网站信息资源的组织构架可分为4个方面。

(1) 类目设置。首先是一级类目设置是否反映网站资源内容的整体分布状况?是否根据用户的对象类型而进行类目设置?

其次是类目划分方法,网站类目的设置是否采取主题和学科双重标准?是否依据用户的使用习惯采用了按学科分类和按主题结合的方式来组织网络信息内容,既有按学科角度设立的文学、艺术、政治、科学与技术、社会科学等大类,也有按主题原则设立的娱乐、休闲、计算机与互联网(电脑与互联网)、新闻、生活服务等大类。这种方法使得主题和学科立类方法进行优势互补,在一定程度上弥补了完全按主题立类无法列举所有主题的缺点,也在一定程度上克服了按学科立类使用性差、用户不易掌握复杂的分类方法的缺点。

再次是类目排列顺序。类目的排列是指同位类目的排列次序,其目的是使类目的排列能体现类目之间的联系,符合人们的思维习惯,并突出常用的重要类目资源。常见的排序方式有三种:一是按系统排序,它是根据类目之间的内在联系排列,列类次序反映客观事物本身

的发展和联系；二是按频率排序，在排序时参考类目被使用的频率把常用的频率高的类目排在最前，频率低的靠后；三是按字顺排列，字母语言可按惯有的字母序列列类，汉语按汉语拼音或笔画、笔顺等字顺排列。网站应根据自身资源特点和用户需求选择最适合的类目排列顺序。

(2) 类目层次。网站层次越多、越深，信息就被组织得越细致，分类体系越严谨，每一个类目下信息的相关性就越高，但是同时也会使底层信息过于专指和隐秘，不容易被查找。因此，类目层次的设置应当适中，过多或过少都是不利的，一般来说，网站类目一般都控制在3~6层，专业栏目有适当延伸是比较恰当的。

(3) 类目注释。用户是通过对比类名及相邻类目来判断浏览的方向和进一步浏览的类目的，这就要求网站通过必要的说明和注释帮助用户了解类目的含义，尽量减少不确定性。目前较常用的内容说明方式有两种：一是通过精练的文字指明该类包含的内容范围和不包含的内容范围；二是用列举下位类的方式揭示本类的内容范围，或提示重点的内容、或提示热点的内容、或提示隐藏较深的内容。

(4) 类目横向关系。任何事物都具有多种属性，具有多向成族的特点，事物之间是相互关联的，构成错综复杂的联系，所以各个类目之间存在着交叉关系。在横向关系揭示上，网络分类体系则采用链接的方式，通过在相关类下重复反映，使其成为类目关系的有机组成部分。目前一些网站对这类关系的揭示，最常见的是将横向关系在相关类下重复反映，并标注表示符。例如，雅虎、搜狐和新浪对于类目间横向关系的揭示都采用链接方式，通过在相关类下重复反映来表明类目间的交叉关系。

3) 链接

链接（包括文字与图形的链接）是网站内部信息之间及网站之间联系的纽带，是网站中用途最广、使用频率最高的一种技术。链接可以使存在于不同服务器上的文件互相连接，使用户能够快捷地从一个网站（或网页）跳到另一个网站（或网页），方便信息的查找和传播。从检索的角度看，信息服务机构组织链接的过程其实就是检索信息的加工和存储的过程；用户通过链接在节点之间浏览的过程其实就是一个查找信息的过程，两者构成了一个完整的检索过程。而 Web 是互联网中基于超媒体的系统，提供多种链接方式，可以全面地表达文件之间的多种联系，包括来源方面的、时间方面的、主题方面的等，为用户提供多条检索途径。

(1) 链接的有效性与合理性。链接的有效性是指链接是否有效？网站的有效链点要求满足以下条件：链点指向非本页面，可以指向其他网站的页面和本网站的其他页面；本站内的一个页面只对应一个有效链点，其他链向本页面的站内链点视为无效；本站链向站外的链点都视为有效链点；网站的默认页面为有效页面；网站的基本单元是网页，除主页外，网页通过链接访问，即一个链点对应一张网页，页内链点将被视为无效链点。是否有重复链接？如果作为链接的页面的基本 URL 和当前页面的 URL 相同，就被认为是重复链接，而不管它的其他属性（有效或者相关）。例如，<http://sim.whu.edu.cn/introduce/>和 <http://sim.whu.edu.cn/introduce/index.php> 就被视为重复链接。由于 URL 中的大小写差异忽略不计，如 <http://sim.whu.edu.cn/introduce/>和 <http://SIM.whu.edu.cn/introduce/> 此类例子也算重复链接。镜像站点和原站点同时链接也被计做重复链接。是否存在死链（即不存在的链接）？死链引发的问题包括：404 错误（指定服务器连接上了，但是 URL 中指定的路径没找到，说明检索到的页面已经被移走或删除了，这通常发生在检索工具的数据库相邻的两次更新之间）和 603 错误（指定服务器没有响应）等。

链接的合理性是指是否链接到资源本身？用户希望最好链接到资源所在的页面而不是该资源所属网站的主页。是否链接到正确的地方？所链接的外部资源是否与网站的主题相关？作为超链接的字串是否过长（如整行、整句都是锚点字串）或过短（如仅用一个字做锚

点)? 施链文字颜色与单纯叙述文字的颜色呈现是否有所不同(如前者用较鲜明抢眼的色彩,后者用较暗、较深的色彩)? 探访过的超链接是否采用低于原超链接亮度的颜色? 一个页面里是否设置过多的链接(否则会影响 Web 页面的流畅与可亲性,一篇文章里提供的文字式链接最好不要超过 10 个以上)? 是否链接到只有两三行注解的页面? 是否链接到尚未完成的页面? 是否链接到无法链接的页面(如出现“文件找不到”(File Not Found)等错误信号)? 是使用相对路径链接还是使用绝对路径链接?(后者的运行效率会更高,移动一组文档时也更容易,相对路径是指相对于当前的工作路径,绝对路径是指一个完整的路径)。但这样又可能引发下文将要论及的侵犯知识产权的问题。为此,需要找到合理地使用链接的方式,实现著作权人、传播者和用户之间的利益均衡,既鼓励作品的创作、合理地组织与丰富网络信息资源,又能充分开发利用现有网络信息。链接是否侵犯了知识产权?

(2) 链接的方便性: 是否提供方便的链接控制功能,使用户可以方便地返回前页、进入下一页、返回首页? 是否在具有前后连续顺序的文件里提供必要的链接,使读者得知自己所在的页面是属于一份较大文件的一部分? 是否同时简明扼要地标明此页、上一页与下一页文件的标题或内容梗概?

(3) 链接提示: 是否有链接提示? 即,当用户把鼠标指向某个链接后,在状态栏是否会出现有关这个链接页面的介绍性信息? 对于文本或图像过大的链接是否有提示? 链接显示出的提示信息是静态文本信息还是动态文本信息? 显示的内容是仅仅局限于文本,还是包含图像、表格、动画、表单等多种页面元素?

4) 导航

对用户友好、完善的导航将引导用户自由浏览文件、信息,提高信息的可用性。但是,网页中的内容是按超文本非线性方式组织的,用户使用这种非线性方式浏览网页内容时,由于超文本的复杂性和受限于用户界面,难以建立起整体结构的概念,可能引起迷失方向和认知负载的问题。因此,导航对于有效展示信息之间的关系及方便用户对信息的利用十分必要。

(1) 导航工具: 导航工具是否有助于用户了解“我在哪里”、“我去过哪里”和“我能去哪里”三个问题? 导航工具能否显示网站的内部结构图? 能否记录并显示用户已走过的轨迹? 通过结构图或轨迹图上的相应节点,是否可以引导用户返回到想去的地方? 在浏览的过程中,是否具有多重选择的导航?

(2) 页面导航的方式: 是否有多种多样的导航方式(线性结构与非线性结构)? 一个网站范围的导航方案是否是一致的? 是否是方便用户使用的?

(3) 导航控制: 是否可以通过分级别制定多重标题,限制用户在某一个主题范围内巡航? 是否能让用户每次点击“返回”标志时均能跳转回上一级节点,不至于在探索过程中迷失方向? 是否提供类似于站点地图的引导方向的工具,以给出知识体系结构的总貌图? 是否在每个页面上端指出用户的浏览路径,每一个层次都可以方便用户的进入? 是否有“上一页”、“下一页”、“回子目录”、“首页”等导航按钮?

5) 检索

建立站内搜索引擎是网站提高信息提供能力的最有效途径。网站是否设置查询文本框,便于站内信息资源甚至站外资源的搜索? 检索方式单一还是多样? 是否既可分类浏览查找,又可输入关键词检索? 是否提供全文检索? 是否提供题名、著者、关键词等多种检索选择? 是否提供布尔逻辑、截词检索等高级检索? 是否可以从文献类型、时间等方面做限定检索? 检索结果是否准确? 对检索的响应速度如何? 是否提供改进检索的建议?

6) 用户友好性

是否有使用指南、导言等联机帮助? 用户是否有不同的语言或字体选择? 是否可以选择文本版本或多媒体版本? 用户是否可以控制信息浏览的顺序(如上页、下页、跳转、末页)?

Web 页面是否便于用户使用? 人机交互功能如何?

3. “新浪”网信息资源组织评析

新浪是目前全球最大规模的中文网站之一, 为世界各地华人提供 Internet 信息服务。在主页的设计与组织结构的设计、链接、导航、检索和用户友好性等方面都做得不错。

1) 主页的设计与安排

新浪的目标用户是全球的中文用户, 在考虑到我国国情的基础上, 新浪在主页上将房产和汽车放在了显眼的位置。网页上的内容繁多, 除了分类目录外, 还有大量的二级子栏目下的详细新闻堆砌在主页上, 用户浏览方便, 主次分明, 图文并茂, 信息量大, 但显得比较凌乱。

这个主页的颜色设计比较统一, 较具美感。但进入主页的时候有很大的动态图像广告, 不止一个, 让人比较反感。除此之外, 网页上与新闻内容相关的图像和动画大小适宜, 设计得体。网页信息组织方式多样化, 分别按主题、产品与服务、用户等方式进行分类。网站主页包含了所有应该具备的必备事项。

2) 组织结构

新浪网的网页组织结构为树状结构与网状结构的结合, 该结构中主页设立若干主要栏目, 每个栏目里的信息再分成一些子栏目, 依此类推。

新浪网用导航系统将网站所有内容组织成一棵由主页生发的树, 用户浏览网站的过程是沿着树上下流动的过程。此外, 在网站中网页之间可以互相连接和跳转, 形成了网状结构。在网状结构中有一个主页, 所有的网页都可以和主页进行链接, 同时, 各个网页之间也可以相互链接。其优点是条理清晰, 访问者可以根据路径清楚地知道自己所在板块的位置, 不会迷路, 并且随时可以到达自己喜欢的页面。其缺点是该结构的浏览效率较低, 链接繁多。

在类目设置上, 新浪的大类基本上反映网络资源内容的整体分布状况, 并根据门户网站用户的特点采用与人们生活密切相关的教育、娱乐、旅行、健康等相关类目的设置方法。它的大类共有 34 类, 包括: 新闻、微博、博客、体育、娱乐、财经、科技、专栏、汽车、视频、房产、读书、女性、乐库、空间、论坛、旅游、游戏等。从中可看出, 新浪的类名比较详细和专指, 这些都是人们生活当中的重要主题。新浪的一级类目都从用户角度出发, 其排列顺序按检索频率确定, 列举高频类、突出热门主题, 如新闻、财经、体育和娱乐。它的二级类目仍按主题出现频率排序而不是字顺排序, 这种频率排序使得类目的排列顺序易于变动, 例如, 在 2012 年伦敦奥运会举行的时候, “奥运”这类就被放在了首位, 便于人们查找。

在类目层次上, 新浪的类目层次不深, 一般为 3~5 层, 这是因为它的类目设置并不是按照严格的上下位间的逻辑关系划分的, 很多概念逻辑是包含关系的上下位类被并列成为同级类目。

在类目注释上, 新浪采用列举下位类的方式揭示本类的内容范围, 或提示重点的内容、或提示热点的内容、或提示隐藏较深的内容, 如“体育: 英超、NBA”。

类目横向关系揭示上, 新浪对于类目间横向关系的揭示都采用链接方式, 通过在相关类下重复反映来表明类目间的交叉关系。但它未采用任何表示符说明类目之间的交叉关系, 用户只能从它的类目路径的变化中发现类目之间的交叉关系。

该网站信息组织体系具有重视以事物为中心设置类目, 类目收录范围广泛, 多重列类、重复反映, 排列方式简便, 直接以语词组织信息, 更新迅速等优点, 但仍存在同位类的设置和排列缺乏必要的规律性、类目归属存在不合理现象和缺乏对知识门类系统显示的能力等问题。

3) 链接

网站有“相关主题”、“相关新闻”等多种链接方式, 施链文字颜色与单纯叙述文字的颜色呈现有所不同, 探访过的超链接采用低于原超链接亮度的颜色, 一个页面里设置的链接较

多。无链接提示,存在无效链接,且在版权声明中标明“不保证为向用户提供便利而设置的外部链接的准确性和完整性”。

4) 导航

新浪网有“网站导航”,每个大类下都列出了下面的二级类目,界面简洁清晰。更难能可贵的是每大类都附有“热门推荐”,其中的主题都是目前关注的热点,为用户快速进入这些热点栏目提供了直接入口。设有“新浪首页”、“新浪导航”、大类首页等按钮,有访问路径提示。但有的页面没有回主页的按钮。

进入每一个具体新闻页面后,在新闻标题的上方都有其类目路径显示,每个类目也都有超链接,如“体育 > 综合体育 > 2013 年羽毛球世锦赛 > 羽毛球 > 正文”。通过页面最上方的一级类目链接也可以方便地跳转到首页和其他栏目。

5) 检索

新浪网既可通过类目层层浏览的形式进行检索,也可输入关键词检索。新浪网采用了自己开发的搜索引擎。

用户可通过新浪搜索引擎将检索范围限定在特定的主题范围内,如新闻、专题、图片、博客、微博、视频、音乐、图片、地图等。而用户在高级检索中可以将资源来源限定在某个大类,如新闻、财经、体育、娱乐等。该搜索引擎支持“AND”、“OR”和“NOT”检索符,支持模糊检索。

6) 用户友好性

网站上有全面的公司与网站介绍、联系方式、问题解答、意见反馈等。用户使用和用户交流都比较方便。

9.2.2 搜索引擎信息资源组织评价

搜索引擎(Search Engine)是指根据一定的策略、运用特定的计算机程序搜集互联网上的信息,在对信息进行组织和处理后,为用户提供检索服务的系统。

1. 搜索引擎的评价标准

搜索引擎类型多样,不同类型的搜索引擎在信息组织方面有自己的特色,各自的评价标准也不尽相同。例如,能否全面地收录某一主题范围的信息资源、是否采用本专业的主题词表是评价专业性搜索引擎的两个重要指标。评价元搜索引擎时则应着重考虑:调用的独立搜索引擎数量的多少,是否允许用户浏览并选择要调用的独立搜索引擎?是否覆盖多种网络资源类型?是否可提供主题范畴的目录服务?是否能够提供统一的检索界面而在不同的引擎中检索?是否支持布尔逻辑检索、短语检索、自然语言检索等高级检索功能?是否能够实现检索请求的“本地化”转换?是否提供了足够多的检索选项和功能?是否可以对来自不同成员引擎的结果去重?是否可以对检索结果进行后处理或重新排序?

由于目前使用最普遍的是综合性独立搜索引擎,且它们是元搜索引擎的重要组成部分,下面仅以这一类搜索引擎的评价标准为讨论对象。综合性的搜索引擎在网络信息资源的组织方面可从以下指标进行评价。

1) 信息采集软件的性能

搜索引擎不是靠人工发现和甄别信息的,而是由一个被称做“蜘蛛”(Spider)等的计算机程序在网中爬行,依据一定的网络协议在互联网中发现、加工、整理信息,并为用户提供检索服务。其数据库由称为“Robots”(或“Spiders”,“Crawler”)的自动检索程序建立,无须人工干预。因而,搜索引擎信息采集软件的性能对搜索引擎信息组织与检索质量的影响非常大,主要考察以下方面。

是否可以模拟人工浏览操作，主动进行信息采集？针对网站内容“漂移”，是否有很强的容错能力（即是否能将类似主题内容的网站汇集在一起）？所采集信息的条目是否具有最小的重复性（优秀的信息采集软件不应当出现同一信息被重复存储的现象）？是否支持任意网页编程方式，包括 ASP、JSP、PHP、CGI、Javascript、VBscript 等？是否可以完整准确地将静态页面和动态页面的数据都采集下来（如网易和 Google 能抓取动态网页）？是否可以支持多种语言编码？是否对各国网站都可以进行信息采集？是否具有良好的负载均衡性能和多线程工作方式，可支持对超过 100 个网站的同时采集？面向行业应用的信息采集软件是否具有良好的扩展性和可配置性以适应不同行业对信息采集的不同需求？

由于搜索引擎的信息采集软件是在后台操作的，对于用户而言是不可见的，但是，它采集的结果可以在数据库的收录范围、数据库容量和更新频率中得以体现。

2) 收录范围与数据库容量

收录范围对于搜索引擎查询结果的质量有直接关系，是评价一个搜索引擎的最基本的指标，主要考察其收录的范围是否完备充分。

(1) 地域范围：大型搜索引擎往往面向全球，收录全球互联网的网站或网页，而且，还专门建立一些地域范围的网站。而一些中小型的搜索引擎则专门收录某一地区的信息，如亚洲地区、亚太地区、我国港台地区，甚至小至我国香港地区。全球性搜索引擎收录的范围虽然广泛，但对某一局部而言，却很可能不如区域性搜索引擎收录某一地区的信息内容丰富和完备。

(2) 语言范围：大型的搜索引擎往往有各种语言版本的网站，各版本的内容也不尽相同。互联网上的中文网站，往往使用不同的语言来编写发表，常用的有 GB（简体中文，大陆地区、新加坡常用）、Big5（繁体中文，我国港台地区常用），其他还有 HZ 码、图形方式等。因此，有些中文搜索引擎主要收录简体中文网站的信息（特别是大陆地区的搜索引擎）；有些中文搜索引擎主要收录繁体中文网站的信息（特别是我国港台地区的搜索引擎）；有些搜索引擎则兼收各种中文网站信息。

(3) 专业范围：综合性的搜索引擎收录各个方面、各个学科、各个行业的信息，组织的范围包括人类知识的各个领域，不论是系统知识还是零散的知识，不论是人文社会科学知识还是自然科学知识，不论是学术研究性知识还是生活服务及娱乐性知识；对搜索的信息进行必要的控制和过滤，防止大量无价值、质量差、死链接或错误链接的信息进入数据库，否则将大大加重用户筛选信息的负担。

(4) 资源类型范围：综合性的搜索引擎应该适合对各种类型信息的组织，包括不同媒体的信息及通过不同的传输协议发布的资源。在互联网资源中，最常见的是 WWW 资源，但也还有其他多种形式的资源，如 FTP、Gopher、BBS、新闻组、博客等。大型搜索引擎往往能够搜索各种类型的资源，它们开辟专门的选项，来搜索 FTP、Gopher、BBS、新闻组、博客中的内容。但有相当多的搜索引擎只收录 WWW 资源，而无法查询其他类型资源。

数据库容量受收录范围的影响，主要指对多少网站（或网页）做索引、收录多少新闻组文章和 FTP 文件等。

3) 数据更新频率

更新周期也能从另一个侧面反映数据库的收录范围，一般来讲，更新快的搜索引擎对最新资料的收录较及时。搜索引擎以多快的频率来更新其数据库是显示其服务质量的重要指标。在不考虑成本的情况下，搜索引擎的数据更新频率当然是越快越好。如果更新频率太慢，跟不上网上信息的更新速度，就会出现错误链接或死链。要满足网上信息动态性的要求，尽量缩小搜索引擎的信息库与网上信息更新的时滞，必须有高智能的自动搜索、分析、标引和著录系统，以最短的搜索周期将变化了的 Web 页的信息加以标引，并追加到数据库中；及时

剔除已成“死链”的链接；建构的知识组织体系和术语系统应具有动态性，以适应网上信息的变化；根据需要增加新的检索手段以满足新类型信息检索的需要。但是，作者在考察中发现，大多数搜索引擎并没有提供数据更新频率方面的资料，只能以最新的问题作为实例获得主观的评价。

4) 检索功能

信息组织的目的是便于检索，提供强劲的检索功能。搜索引擎是否提供既能满足一般用户的简单检索，又能提供专指检索（高级检索）。高级检索是否为图形界面？是否带有选项功能的下拉菜单？查询方式是否可以按用户所需的方式进行设定？中文搜索引擎还要看其能否自动识别中英文、能否自动进行内码转换、能否自动进行中文分词？提供了表 9.1 中的哪些检索功能？

当然，不是每一个搜索引擎都必须同时具备上述功能，但至少应该支持布尔逻辑等基本检索，而大型搜索引擎的检索功能应该更多一些。

表 9.1 搜索引擎的常用检索功能

检 索 功 能	英 文 名 称	备 注
布尔逻辑检索	Boolean Logic Search	有的用 AND, OR, NOT (或者小写), 有的以符号 +、-、* 代替, 还有的直接把布尔逻辑算符隐含在菜单中
短语检索	Phrase Search	又称“精确检索”, 检索出与“ ”内形式完全相同的短语
截词检索	Truncation/Wildcats	允许在检索标识中保留相同的部分, 用相应的截词符代替可变化部分, 以扩大检索范围
词根检索	Stemming	可将搜索关键词进行延伸, 查找词根相同的记录
邻近检索	Proximity Search	又称位置算符检索, 用一些特定的位置算符表达检索词与检索词之间的位置关系, 如 (W), (N), InfoSeek 的 far, before, 目前具有这一功能的搜索引擎很少
区分大小写的检索	Case-sensitive Search	如果用户输入时用小写字母表示, 既匹配大写又匹配小写, 如输入 china, 将检索出 china、China 和 CHINA; 但如果用大写字母表示, 搜索工具认为用户指定了只要大写, 就只会查找那些与用户键入的输入形式完全相同的结果
全文检索	Full-text Search	对网页全文中的每个词进行检索。为了提高检索效率, 英文搜索引擎常常将一些介词、冠词等作为禁用词。仅仅用这些词来进行检索, 搜索引擎将不予作答
模糊检索	Fuzzy Search	又称概念检索, 搜索引擎不仅反馈包括了关键词的网页, 同时也发来与关键词意义相近的内容
自然语言检索	Natural language Search	又称智能检索, 直接采用自然语言中的字、词、句作为提问式进行检索
多语种检索	Multilingual Search	提供多种语言的检索环境供用户选择, 系统按用户选定的语种进行检索并反馈结果
按范例查询	Query-By-Example Search	指示搜索引擎对与某个具体文档相类似的文档进行搜索, 也称为“相近搜索结果”
限制检索	Limit Search	将检索限定在某一范围中, 包括类别范围、地域范围、时间范围、语言范围、网站类型、文件类型、域名、位置等。限制实现的方法各不相同, 有的在关键词前后加特殊的命令实现, 如 title、link、url、.com、image, 有些通过下拉菜单实现

5) 检索效果

对于检索效果的评价主要从以下几个方面进行。

(1) 响应时间。

从搜索引擎的实用性来看,必须保证对用户检索表达式一定的响应速度,在这个基础上才谈得上使用的方便性等其他因素。有的搜索引擎对每一次检索都给出了搜索用时(如 Google),这为用户对响应时间的比较提供了很好的依据,无须用户来计算时间,希望更多的工具提供这种功能

但在网络环境下,响应时间不仅取决于检索工具本身的响应速度,还在相当大的程度上取决于用户使用的通信设备、网络的拥挤程度等外部因素。因此,在计算响应时间时,应该在相同的时间,在相同的软/硬件环境下,对同一个检索课题(由于不同的工具支持的运算符与检索式的构成规则不同,检索式不可能统一要求)的响应情况进行量化评价。

(2) 查全率。

查准率和查全率一直是评价信息检索系统检索性能的最重要的指标。但真实查全率在网络环境下也很难获得。真实的查全率,即检索出的相关文献量和文献空间中所有相关文献量的比率。网络信息数量庞大且瞬息万变,不同搜索引擎标引的深度不同,用户不愿意也没有足够的时间在太多的结果里筛选适合自己的资源。因此,查全率对于网络信息检索工具的评价无法操作也不具有现实意义。

(3) 查准率。

与查全率相同,真实的查准率,即检索出相关文献的数量和检索出的文献总量的比率,也是很难计算的。因为对于命中结果数太大的检索课题来说,相关性判断的工作量相当大,要对它们进行一一判断是不可能的。但对用户来说,在网络信息资源数量如此之大的今天,查准率远比查全率有意义。目前比较成功的一种计算查准率的替代方法是两位美国研究人员 H.Vernon Leighton 和 Jaideep Srivastava 提出的“相关性范畴”概念和“前 X 命中记录查准率”,但该计算太复杂。全球搜索范围最大的 Google 认为,大多数用户均可在前 10 个结果中找到所需资料,所以 Google 预先设定为 10 项。

笔者在结合二者算法的基础上认为,应将检准率的范围限定为前 30 个结果中符合要求的比率。对每个检索结果进行相关度 3 级评分,相关则打 1 分,比较相关为 0.5 分,不相关则为 0 分。重复出现的网页只记分一次。笔者将每个网页的相关得分之和除以 30,乘以 100%。

(4) 链接的可靠性。

这是网络信息检索工具性能评价特有的指标。这个指标与数据更新频率和查准率有关。链接是否可靠?有无断链、死链现象(断链、死链会造成找不到原始文献的情况出现,那么命中记录再多也没有用)?

6) 检索结果的显示

搜索引擎总是要将检索结果返回给用户,而结果显示的好坏直接影响到搜索引擎的使用效果。

(1) 输出格式的灵活性。

不同的用户有不同的需求,检索工具能否提供多种检索结果的输出格式供用户选择?

(2) 显示的内容。

显示的结果越详细,就越有利于用户不看全文即可决定对所检信息的取舍吗?只是简单地显示文件标题与 URL 可以吗?是否更详细地显示包括检索式、搜索用时、关键词、原网页所用的内码、文件大小、文件日期、内容摘要或评价等内容?是否提供到文献全文的链接?

显示的摘要是提纲挈领式的还是简单地摘录原网站中的部分内容？前者的信息含量要高得多。一般搜索引擎的网站提要都是各网站向搜索引擎注册时自己提供的。好的搜索引擎，要对来注册的提要逐一检查核实，增删修改，这往往需要很大的工作量。有些搜索引擎为减轻工作量而使用自动注册，或不进行核实，这样其内容提要就会存在不准确或注册者的自我夸张的情况。理想的内容提要应该是搜索引擎自己的工作人编写的，并提供较权威的评价。可惜，这样的搜索引擎实在是凤毛麟角。

(3) 显示数量。

每页显示的记录数是否可以由用户选择或设定？

(4) 检索结果的排序。

检索结果的排序依据是什么？是否依命中相关度（Relevance）排列检索结果？是否提供逆时序排列（Reverse Chronological）的选择？或者依概念、网址、域名、声望和链接等标准排列？是否提供多种排序方法可供用户选择？

(5) 对搜索结果的处理。

是否具有对搜索结果的去重功能（一个网站最多只能有一页出现在排名靠前的搜索结果中）？是否可进一步进行限定检索？是否能够删除重复的链接以保证在搜索结果中用户可以有更多更好的选择机会？能否保存检索式并对其修改？可否在返回的结果中进行二次检索？能否将用户的检索结果组织到不同的文件夹中（如 Northern Light 可以对用户不同的检索结果提出各种备选的文件名，供用户选择）？

7) 用户友好性

搜索引擎的前端界面本身就是一个网站，其页面组织与用户友好性的评价与“网站方式”基本相同。同时，搜索引擎最主要的目的就是尽可能完美地服务于用户的信息检索，简单、易用应该是检索工具永恒不变的追求。检索过程是以完善和改变用户知识结构为目的的过程，提高搜索引擎的智能化程度也是提高易用性的重要方面。

搜索引擎是否在学科领域知识和语言知识方面对用户予以充分的支持？是否具有扩展通向相关信息路径的支持？这样，可以使用户通过与系统的交互作用及磨合，逐步缩小信息表述的差距，将用户的认知负担降到最低。不同用户利用网络信息有相对稳定的一面，搜索引擎是否为不同用户建立自己的知识分类系统提供相关的功能？页面是否能根据用户需求进行动态定制？是否有丰富的联机帮助文件和提供动态的帮助？是否提供丰富的帮助内容是其页面组织与用户界面友好性评价中一项必不可少的指标。用户是否可以自定义检索命令？是否有菜单引导、表格引导、屏幕提示、位置提示和出错告警等辅助手段？是否提供许多热键？是否支持触摸式输入和语音输入？

8) 其他功能与服务

是否提供新闻、天气预报、旅游、黄页、电话号码、航班和列车时刻表、地图等常用信息？搜索引擎，尤其是国外的搜索引擎，为了吸引用户，在提供搜索服务之外，还提供其他相关服务，从而获取更多的广告收益。

是否增加特色服务？网上检索工具已不仅仅是单纯的检索工具，正在向其他服务范畴扩展，提供免费邮箱、自动翻译、网上聊天、站点评论、新闻报道、股票点评等，以多种形式满足用户的需要。

是否针对不同地区的用户提供本地化服务？随着上网用户的不断增加，信道越来越拥挤，远程终端上网速度越来越慢，搜索引擎提供的本地化服务可以分流用户，提高上网查询速度。Google 和 Bing 都在世界各地设立了分支机构。

是否提供多媒体信息的检索？

2. 著名综合搜索引擎 Google 的评析

美国知名的数字媒体评估公司 ComScore 统计数据显示^①, 在 2012 年 11 月和 12 月份期间, 全球用户通过 Google 进行的搜索查询达到了 1147 亿次, 市场份额为 65.2%, 市场占有率远远高于其他搜索引擎, 特别是在欧洲、北美和拉丁美洲这些以英语为母语的国家。可以说, Google 应该是在全球范围内影响最大的搜索引擎。下面根据以上提出的搜索引擎评价指标对 Google 进行评析。

1) 信息采集软件的性能

Google 储存网页的快照, 当存有网页的服务器暂时出现故障时用户仍可浏览该网页的内容。如果找不到服务器, Google 储存的网页快照也可救急。虽然网页快照中的信息可能不是最新的, 但在网页快照中查找资料要比在实际网页中快得多。

搜索引擎的信息采集软件在后台操作, 我们无法一一详细评价, 但可从其数据库的收录范围、数据库容量和更新频率中来分析信息采集软件的性能。

2) 收录范围与数据库容量

Google 作为全球最大的搜索引擎, Googlebot 网页爬虫每天都会走过大约 200 亿个网页, 其目录中收录全球范围内的网页总量达 30 万亿个以上。而且借助和 America Online 等公司和出版社、图书馆的合作, 其数据库容量远远大于其他搜索引擎, 内容广泛, 语种丰富。它可搜索的文件类型有: 普通的网页 (html)、Adobe 可移植文档格式 (pdf)、Microsoft Excel (xls)、Microsoft Power Point (ppt)、Microsoft Word (doc)、Rich Text Format (rtf) 和 Shockwave Flash (swf) 等 14 种。

3) 数据更新频率

Google 研究者们发现了一个很简单的方法测算 Google 的更新频率, 就是查看 Google 的几个镜像站点是否相同, 因为在 Google 更新期间, 它总是将 www2.google.com 或者 www3.google.com 作为更新测试站点, 在这期间, 后两个站点中索引的页面数量和主站点将会不同。三个网站搜索同一个关键词, 得到的结果的数量不相同, 这就说明 Google 正在更新。Google 这种大规模的更新在国外被称为 Google Dance (狗狗跳舞), 网络上已经出现很多专门协助用户检查 Google 是否正在跳舞的工具 (如: <http://www.searchbliss.com/seo-tools/google-dance-tool.asp>), 通过这个工具, 用户只需输入一个关键词然后点击一个按钮, 就可以看到 Google 三个镜像站点的搜索结果, 很方便对比。

数据库何时更新, Google 并没有严格的规定, 但根据 Google 更新周期的一般规律来说, 平均的小更新时间为每 7 天一次, 大更新则为每月一次, 前者对网站排名影响小, 后者对网站排名影响大。网页排名 (PR, Page Rank) 值更新完一次需要 3 个月, 网站的 PR 值是 Google 搜索排名算法中的一个组成部分, 级别从 1~10 级, 10 级为满分, PR 值越高说明该网页在搜索排名中的地位越重要, 也就是说, 在其他条件相同的情况下, PR 值高的网站在 Google 搜索结果的排名中有优先权。总体而言, Google 的更新频率比较高, 其中“谷歌趋势” (Google Trend) 可做到每天更新, 它通过分析全球数以十亿计的搜索结果, 告诉用户某一搜索关键词在 Google 被搜索的频率和相关统计数据。

4) 检索功能

Google 的检索功能非常强大, 不仅可提供多种人性化的检索模式、支持多种检索符, 还针对各种常用资源类型设置了专门的检索服务。

^① Yandex Just Passed Bing to Become 4th Largest Global Search Engine. <http://searchenginewatch.com/article/2242374/Yandex-Just-Passed-Bing-to-Become-4th-Largest-Global-Search-Engine> (访问时间: 2013-08-10)

Google 提供的检索模式有两种：简单检索和高级检索。用户若熟悉各种 Google 的检索式和检索符，可直接在简单检索的文本框中输入相关检索式进行检索。高级检索是利用带有选项功能的下拉菜单来提供方便操作的检索服务。这些选项功能包括关键词限定、网页语言的种类、网页区域（属于哪个国家）、文件格式（可将检索结果限定在 pdf、xls、doc、ppt 等中的任意一种）、日期（网页发布的时间）、字词位置（可将关键词限定在网页的标题、内文、网页内的网址或网页的链接内）、网域（在某个特定的域或站点中进行搜索）、使用权限（是否可随意使用、共享或修改，能否用于商业目的）、类似网页（搜索类似检索文本框中所输网页的网页）和链接（搜索与检索文本框中所输网页存在链接的网页）。

Google 支持的检索符有布尔逻辑运算符“或”（OR，-）、“非”（-，NOT），且自动在关键词之间添加“AND”，不需要用户特别输入；支持短语检索（“”），可作多种限定检索，并支持一些特殊的命令，如 site:、all in title:、date:、define:、movies: 等。

Google 采用主题目录和搜索引擎相结合的方式为用户提供专门检索服务。除了关键词检索外，用户还可以通过分类列表浏览网站导航和“热榜”（各类信息资源中检索频率最高的关键词集合）来进行网页浏览。用户按主题目录类别搜索的方式可以根据主题来缩小搜索范围。例如，在 Google 目录的 Science>Astronomy 类别中搜索“Saturn”，可以找到只与 Saturn（土星）有关的信息，而不会找到“Saturn”牌汽车、“Saturn”游戏系统或“Saturn”的其他含义。

5) 检索效果

Google 在如此大的库容量情况下仍能保持非常快的响应速率，并且对每一次检索都给出了搜索用时，一目了然。

Google 链接的可靠性也非常高，笔者在使用 Google 的过程中很少发现有断链、死链现象出现。

6) 检索结果的显示

Google 的检索结果输出格式统一，用户不能自行选择输出格式。其检索结果显示的内容包括检索式、搜索用时、命中记录总数、源网页的网站名称、网站链接和地址、内容摘要、网页大小、网页快照和类似网页。对于搜索结果中大型的网站，还在检索结果中列有该网站的热点内容关键词，用户可直接点击这些关键词进行查看。Google 提供的检索结果显示项虽然丰富，但其网页内容摘要只是简单地摘录原网站中的部分内容，并没有原创性的评价。

Google 检索结果每页显示的记录数量可由用户选择，分别为 10 项、20 项、30 项、50 项或 100 项，用户也可在“使用偏好”中设定每页显示检索结果数量。检索结果的排序主要依照相关度，用户不能自行选择检索结果排序方法。其搜索结果重组功能使一个网站中不可能有两页出现在同一搜索结果中，而该网站的第 2 页一般出现在搜索结果列表第 1 页下面，如果要查看这一网站中的其他网页，可以点击“More results”。

7) 用户友好性

Google 的界面非常简洁易用，页面上只有醒目的检索框和简单的主题导航栏。其智能化程度也较高，在用户检索过程中，当关键词出错时，Google 还会给出查询建议供用户参考，非常有用。并根据用户的不同需求开发了相应的检索工具，如学术搜索、生活搜索、地图搜索、博客搜索、视频搜索、图片搜索等，而且，这些功能分别提供简单检索与高级检索选项。

不同用户还可根据自身喜好对 Google 进行动态设置，建立符合自身检索习惯的搜索引擎，包括界面语言设置、搜索语言设置、结果数量设置、结果视窗设置、中文简繁转换设置和查询建议设置等。在英文版 Google 中，用户还通过 iGoogle 将自己感兴趣的新闻、游戏等内容添加到 Google 的主页上以方便浏览。

Google 的帮助内容也比较丰富，它将 Google 所有的产品分为搜索、购物和浏览网页等

6 大类，用户可根据类别选择有关产品的帮助信息。除了网站本身提供的帮助信息外，帮助页面还设有向其他用户请教的模块链接以实现在线实时咨询的功能。

8) 其他功能与服务

Google 的产品非常丰富，中文版 Google 提供的产品服务共有 31 种，而英文版的服务高达 43 种。这些服务类型多样，既有各类型资源的搜索服务，也有 Google 开发的一些常用工具软件、移动通信服务和交流工具。

Google 的专项资源搜索服务非常细化，包括博客搜索、Google 财经、地图搜索、热门榜单关键词搜索、生活搜索、视频搜索、图片搜索、网页目录浏览、学术搜索、新闻搜索、美国专利全文搜索等。

Google 的特色服务种类繁多，实用易用是其最大的特点。如根据财经信息需求量大的特点，开辟了财经板块，其中包括商业信息、财经新闻、实时股价和动态图图表，内容实时更新。用户还可通过“快讯”定制实时新闻，直接从邮箱中阅读。

Google 开发的一些常用工具软件有 Google 工具栏（为用户的浏览器配置搜索框以便随时搜索）、Picasa 图片管理软件（查找、编辑和管理计算机上所有图片）、在线翻译语言工具、Google 文档（创建在线文档、电子表格和演示文稿以实现实时共享和协作）、SketchUp（3D 绘图软件）、Google 拼音输入法、日历管理等。在用户交流方面，Google 不仅提供免费邮箱和论坛，在英文版中还提供实时聊天工具“Talk”和博客创建工具“Blogger”。

Google 英文版所提供的服务比中文版要多得多，一些非常有特色的服务值得一提，如 Product Search 专门提供商品信息比较的服务，Google Health 是供用户在线组织个人医疗记录信息的工具。Google 若能将这些特色服务根据我国特色汉化和本土化，方便我国用户使用，将会大大扩展其用户群体。

在多媒体信息检索方面，Google 有专门的图片和视频搜索服务，并收购了全球最大的视频网站 Youtube，这都充分说明了 Google 在多媒体信息检索方面做出的努力。

综上所述，Google 作为全球最具影响的搜索引擎，各方面做得都比较出色，但也存在一些问题。例如面对各国本土搜索引擎的挑战而自身本土化问题较难解决及多媒体信息检索仍然无法完全突破等问题。

9.2.3 学科信息门户资源组织评价

学科信息门户（Subject Information Gateway，简称 SIG）将特定学科领域的信息资源、工具和服务集成一个整体，为用户提供方便和统一的检索和服务入口。作为网络信息服务平台，SIG 通过灵活的资源整合、可靠的信息组织，无缝地链接用户所需的信息资源和信息服务，将一个分布式的、繁杂的网络信息空间组织成一个方便的用户信息系统，提供浏览和检索双重功能来满足用户科学研究和教育等方面的信息需求，并在此基础上支持个性化定制服务。

1. 学科信息门户资源组织特点

学科信息门户按照某学科（专题）用户的信息需求对网络中相关的资源进行针对性、深入的揭示，在为用户“导航”的同时提供专门的信息检索服务，有助于专业用户在本领域的“信息超市”（Information Supermarket）中实现“一站式检索”（One-site Search）和获得高质量的资源。对图书馆来说，学科信息门户拓宽了“馆藏”，而对整个网络而言，其信息的序化程度得到提高。学科信息门户资源组织呈现出以下一些特点：

1) 信息资源内容的专业性

搜索引擎是目前互联网上使用率最高的信息组织工具，但由于其没有采用受控语言，在

学术信息的检索方面,检准率较低,并且难以检索出隐性信息。相对于搜索引擎而言,学科信息门户根据特定主题收集信息资源,针对特定的专业领域满足特定用户需求,具有一定的优势。此外,门户的信息组织还经过图书情报学、计算机科学和相关学科专家的共同参与,保证了信息内容的高质量和专业性。

2) 信息资源与应用的集成整合

学科信息门户致力于将特定学科领域的信息资源、工具与服务集成到一个整体中,为用户提供一个方便的信息检索和服务入口。学科信息门户提供的每一种信息资源都经过图书馆员和学科专家的选择与描述,收录的资源包含电子期刊、数字化图书、报告、论文、书目、教育软件、多媒体资料电子新闻稿和重要的学术机构的网站等。学科信息门户资源需经过深层次组织加工,根据对资源知识内容及关系的分析,揭示和反映资源内在知识性内容,形成高质量、集成性的信息服务体系。并且,这些信息与服务有机地集成在一个统一的界面中。

3) 跨系统一站式检索

学科信息门户通过共同的表达和一致的用户界面,使其更易于使用。用户在一个搜索界面,将搜索请求一次性输入就可实现对学科信息门户多种资源和数据库的信息查询。学科信息门户还将各个系统的检索结果汇集起来,以统一的界面展示给用户,使用户的搜索方便而高效。由于界面统一并遵循用户习惯,用户无需进行培训就能方便地发现和搜索到所需信息。

4) 个性化信息服务

学科信息门户从一站式的以信息资源互动为中心的网站服务,向一体化的以用户行为、需求、体验为中心的网络服务系统转变。学科信息门户可根据用户不同的角色预设不同界面内容,可基于用户所属的角色提供给用户相应的内容。根据用户需求与偏好的描述信息,或通过用户信息访问行为的动态分析来推测用户意图,进行信息过滤和信息推荐,对不同用户提供不同的内容和用户界面。

2. 学科信息门户资源组织评价标准

为了优化现有的学科信息门户,并为新的学科信息门户建设提供参考,需要对门户的信息组织提出一套评价指标。学科信息门户在网络中是以网站的形式呈现的,因此 9.2.1 节中关于网站信息资源组织评价的标准在这里同样适用。同时,根据门户资源组织的特点,在组织信息中还应该尽可能全面收录本学科资源、严格选择资源、对资源进行高质量的元数据描述、构建合理的分类体系和尽量运用受控词表、重视互操作性、运用相关技术、定期更新和维护、提供个性化服务等。具体评价指标如下:

1) 学科资源的覆盖率

判断一个学科信息门户的优劣,首要是看其所指引的有关资源是否尽可能地涵盖本学科(主题)领域重要的资源。是否汇集了某学科(专题)领域重要的相关信息资源(涉及资源的内容、时间跨度、语种与地域范围等)?门户中收录的信息资源类型是否多样?是否能方便用户对某一学科(专题)信息资源的“一站式检索”?如学科信息门户 INFOMINE 的记录达到了 12 万余条,涵盖了数据库、电子期刊、电子图书、电子公告板、邮件列表、在线图书馆馆藏目录、论文、研究人员目录等诸多类型的信息资源,并在主页上提供了一站式检索接口和高级检索。

理想的学科信息门户不仅应收录互联网资源而且还要囊括图书馆馆藏实体资源(包括二次文献数据库、全文数据库、馆藏目录、联合目录等),成为集成化的信息资源系统,两种资源在同一界面实现无缝存取(Seamless Access),整合成便于检索和使用的有机整体,即复合图书馆(Hybrid Libraries)形式。

2) 资源的质量

学科信息门户资源是为用户提供信息服务的根本,因此资源质量高低是判断学科信息门户价值最主要的标准。门户的资源选择必须有一套符合该学科特点、既定用户需求、服务宗旨、规模及经费支持等方面要求的资源选择标准。

门户信息资源在内容上既要注意信息的准确性、权威性、客观性,还要尽量保证信息的唯一性、新颖性、完整性与针对性,所指引的网址要尽量接近主题内容,尽量减少用户点击的次数。指引的资源是否适合目标对象?对资源的描述是否准确?指向的资源地址是否正确?是否链接资源的原始网址以使用户访问到最及时和最权威的信息?在形式上要考察信息的格式,看是否是标准通用的或常见的格式?用户使用的便利性如何(考虑用户访问本网站采集的资源时所需的硬件、软件和连接方式等要求)?导航是否明确、清晰?排版结构的美观程度以及资源的可存取性与可用性如何(注意被链接资源的注册要求、收费规则、知识产权声明与特殊服务规则等;优先选择网上免费资源,当网络用户与镜像站点的“电子距离”比原始站点更近时,要链接镜像站点以便于用户进行更有效的存取)?

3) 资源的元数据描述

对选取的信息资源进行高质量的元数据描述是为用户提供优质信息服务的前提。元数据是关于数据的数据,它对信息资源或数据对象进行描述,目的在于使用户能够发现、识别、评价资源,并对相关的信息资源进行选择、定位和利用,追踪信息资源在使用过程中的变化,实现信息资源的整合、有效管理和长期保存。

对资源的描述是否采用国际通用的元数据、标记语言、分类法与词表?学科信息门户对资源的描述应该优先采用国际通用的元数据与标记语言,元数据元素中的“主题”描述要利用国内外著名的分类法与主题词表。例如,澳大利亚要求学科信息门户全部采用已有的元数据标准[包括都柏林核心元数据(DC)、MARC和澳大利亚政府查找服务(Australian Government Locator Service,简称AGLS)等元数据]并对自己独有的元数据元素进行标引,用以支持互联网内容选择平台(Platform for Internet Content Selection,简称PICS)。INFOMINE的记录以HTML语言表示,主题标引使用《国会图书馆主题词表》(Library of Congress Subject Heading,简称LCSH),记录还可以转换成MARC格式。

4) 分类体系和受控词表

分类体系是对学科信息门户收集的资源实施分类组织以及用户进行信息资源浏览与检索的依据与桥梁,其科学性十分重要。学科信息门户可以采用已有的文献分类法(包括综合性的分类法与专业或专题分类法),也可以结合自身特点与目的对已有的分类法进行适当改进,或者采用自编的分类法。在分类体系的构建中,分类表的展现应力求简单、明了,尽可能将所有的一级类目展现在一个页面;还要充分利用网络方便的超链接功能,对具有多重隶属关系与相关关系的类目设置合理的参照系统。学科信息门户词表的设置与否对于本专业的深入检索影响很大。学科信息门户应建立符合相关专业特点的主题词表和分类体系。表9.2列出了国内外部分学科信息门户及其采用的分类法和受控词表。

表 9.2 国内外部分学科信息门户及其采用的分类法和受控词表

国家	名称	学科	开发者	URL	分类法	受控词表
美国	The WWW Virtual Library	综合	Tim Berners-Lee 与自愿者	http://vlib.org (拥有英语、法语、西班牙语、中文版本)	自编	无
美国	INFOMINE	综合	加州大学河滨分校图书馆	http://infomine.ucr.edu	LCC	LCSH

续表

国家	名称	学科	开发者	URL	分类法	受控词表
英国	Intute	综合	JISC、AHRC	http://www.intute.ac.uk	UDC、DDC、LCC 等	CAB、AAT、HASSET 等
美国	The Gateway to 21st Century Skills	教育	教育部、国家教育图书馆等	http://www.thegateway.org	自编	ERIC
英国	The Arts and Humanities Data Service (AHDS)	艺术与人文	JISC 等资助	http://www.ahds.ac.uk	自编	AHDS Subjects
美国	IPL2	综合	德雷塞尔大学等	http://www.ipl2.org	自编	LCSH
中国	中国科学院国家科学图书馆学科信息门户系列	数理化、生物、资源环境、图书情报	中国科学院	http://www.las.ac.cn	中国图书馆分类法	中国分类主题词表

注：LCC 全称为 Library of Congress Classification；UDC 全称为 Universal Decimal Classification；DDC 全称为 Dewey Decimal Classification；LCSH 全称为 The Library of Congress Subject Headings；ERIC 全称为 Education Resources Information Center Description；CAB 全称为 CAB International Agriculture Thesaurus；AAT 全称为 Art & Architecture Thesaurus；HASSET 全称为 Humanities and Social Electronic Thesaurus。

5) 互操作性

学科信息门户中各信息源数据库与信息平台差异可能很大，为了在统一的界面使用来源各异的网络资源，学科信息门户必须具有异构计算机软硬件平台间良好的互操作性，具有跨门户检索的能力。

为了提高互操作性，美国国家自然科学基金会（NSF）曾资助的 Isaac Network 项目采用 DC 作为元数据，以 Linux 作为平台，以 Lightweight Directory Access Protocol (LDAP) 作为信息查询与交换软件，而以代号为 RFC2651 的通用索引构建协议（The Architecture of the Common Indexing Protocol，简称 CIP）作为索引编制与互换协议。通过该门户，用户可以对 SOSIG、BUBL LINK、EEVL、EdNA、MathGuide、GeoGuide、OMNI 等近 20 个学科信息门户进行跨门户的检索。同样由 NSF 资助的“全国科学、数学、工程和技术教育数字图书馆”（SMETE）项目则将多个分布式学科信息门户作为整个数字信息资源的整合机制和服务渠道，允许用户通过该门户体系检索和调用各种不同的信息资源与服务。

6) 相关技术的应用

学科信息门户的运行仅靠图书馆的理念是不够的，还要涉及大量的信息技术，如虚拟现实（Virtual Reality，简称 VR）技术、虚拟专用网（Virtual Private Networks，简称 VPN）技术、虚拟局域网（Virtual Local Area Network，简称 VLAN）技术、虚拟数据库（Virtual Database，简称 VDB）技术、通用对象请求代理体系结构（Common Object Request Broker Architecture，简称 CORBA）技术等。例如，利用 VPN 可以解决学科信息门户信息共享的安全问题。信息推送（Push）技术通过信息代理机制，在用户初次使用时设定所需的信息后，能通过推送（Push）或网播（Net-casting）的方式把网上相关信息送到用户面前。这种基于 Push 技术的 Internet 信息检索技术既为用户搜索、浏览网上的相关信息提供了快捷入口，又为学科信息门户在广域网内的信息共享提供了技术支持。

新技术的应用也是评价学科信息门户的标准之一。随着 Web2.0 相关技术和应用在互联

网中的兴起,很多学科信息门户也应用了相关的技术,如在学科信息门户中开设了博客、提供了 RSS 推送订阅、设置了社会网络分享工具等。

7) 个性化与人性化服务

网络环境下,用户需求的变化除了需求量上的增长外,还表现为信息需求复杂程度的提高,包括:用户构成逐渐多样化、复杂化,不同年龄、性别、文化程度、国别、信仰的人士有着不同的信息需求。同一个用户在学习、娱乐、工作等不同的活动中也有着不同的信息需求,他们希望有一个系统能直接、深入、有效地支持其检索、处理信息和利用信息来解决问题,帮助其建立个人的信息空间。用户信息需求的个性化要求学科信息门户在提供信息浏览与检索等基本服务的同时,还要利用网络新技术,跟踪用户需求,主动地为用户提供新资源通报、信息推送与定制服务。学科信息门户还必须利用可视化等技术增强用户界面的友好性,注重帮助功能的提供,体现对用户的人文关怀,注意尊重与保护合法用户的权利与个人隐私。

8) 更新与维护

学科信息门户的更新与维护包括三个方面:

一是信息资源的更新与添加。由于网络资源站点的增加与频繁更改,学科信息门户要真正成为互联网信息的深层次组织工具,必须及时更新其收录的资源,保证信息资源的高质量。学科信息门户要改变因追求高质量而过于依靠人工参与的状况,要不断提高智能化水平,充分利用网上自动漫游、自动跟踪、自动分类和自动标引技术,采用人机结合的工作方式,为用户提供更优质、高效的服务,还可以鼓励用户推荐资源或参与门户的维护。

二是学科信息门户中信息资源的安全。数字化方式存储的信息,极易受到计算机病毒、黑客入侵等的干扰和破坏,资源的安全受到了许多的挑战。学科信息门户的安全性问题应该受到重视,在门户的日常管理中要注意对资源进行备份与保存。

三是学科信息门户服务器持续服务的保障。用户在使用学科新门户过程中,如经常遇到服务器不能连接或有的功能失效等情况,就会失去对该学科信息门户的信心。因此,需要学科信息门户的管理人员提供 7×24 小时维护,确保服务不间断。

判断学科信息门户更新与维护的标准包括:更新是否及时?是否有专门的工作人员不断追加新的网络资源?是否及时剔除错链、死链?整个系统在结构上是否为一个活的系统?是否可根据用户需求和网络信息资源的变化及时对词表和分类体系进行调整?维护是否比较容易?软件是否可靠?是否对资源进行日常性的备份与保存?

Intute 整合了 Altis、Artifact、BIOME、EEVL、GSource、Humbul、PSIgate、SOSIG 等非常有名的学科信息资源门户,向用户提供网络资源发现服务。但 Intute 已于 2011 年 7 月停止了信息资源的更新,随着时间的推移,它的信息提供功能和学术价值将逐渐失去。

3. 学科信息门户实例 IPL2 评价

IPL2 项目由原著名的“图书馆员互联网索引”(Librarians' Index to the Internet, 简称 LII)和“互联网公共图书馆”(Internet Public Library, 简称 IPL)于 2010 年 1 月合并升级而来,它依靠图书情报从业人员和图书情报学院(系)的师生参与优质网络信息的采集、选择、评价和重组,提供用户浏览和检索可信赖的网络资源的通道,成为图书情报人员参与网络信息组织的典范。下面根据以上提出的学科信息门户评价指标对 IPL2 进行评析。

1) 学科资源的覆盖率

LII 始于加州大学伯克利分校参考馆员 Carole. Leta 1990 年制作的 Gopher 书签文档,经过 20 年的发展,成为一个可供检索的、提要性的学科信息资源目录。IPL 是由美国密歇根大学信息管理学院的教师及该校的图书馆员们于 1995 年开发,经过 10 多年发展,IPL 项目即已发展成为了一个“主要以虚拟服务为特征”的数字化图书馆。新建立的 IPL2 综合了 LII

和 IPL 的资源和服务,毫无疑问, IPL2 仍然是一个提供“一站式检索”的学科信息门户。

2) 资源的质量

IPL2 对收录信息资源的选择有着严格的控制标准,首先确定了按照用户需求选择资源的一般原则,所选资源以英语为主,但不排除其他语种。IPL2 还制定了具体的资源选择标准,包括:所选信息资源的可用性、权威性、合法性;对信息资源网站的设计及功能也提出了标准,如网站设计要清晰、井井有条,网站的主要提供功能(如检索)应该是可用的,网站不提示错误信息等;对信息资源本身也提出了相关标准,如网站内容应该很少出现拼写的错误,网站应该有实际的内容,网站应该在相关学科领域进行及时的更新等;列举了 IPL2 不收录的资源,如资源违反美国版权法、没有实质内容的商业网站、存在极端主义观点的网站、不提供免费内容的付费网站等;为儿童和青少年提供的信息资源,必须支持他们的教育目标,不能含有亵渎、色情或暴力的材料等。此外, IPL2 还对收录报纸杂志、Web 目录、搜索引擎、网络应用、博客提出了标准。并且,在以上标准的严格要求下,每一种信息资源都经过图书馆员对其价值进行评价。

3) 资源的元数据描述

IPL2 资源的每条记录都给出了详细的介绍,包括资源名称、URL、摘要、主题词、记录创始人与时间,以及记录修改人与时间等。摘要由 IPL2 工作人员撰写,使读者在进入一个网站前便可了解其主要内容。IPL2 链接的基本是资源的原始地址。对工作流程、著录项目与规则以及注意事项均做了非常具体的规定,因而能够保证资源描述的质量与一致性。

4) 分类体系、受控词表与互操作性

IPL2 向用户提供了一个简洁、易于使用的界面,用户可通过浏览和检索获取所需信息资源,网站主页上提供了“聚焦资源”(Spotlight)和“特色精选资源”(Featured)的推荐。资源浏览部分为:按主题资源(Resources by Subject)、报纸与杂志(Newspaper & Magazines)、IPL2 特藏(Special Collections Created by ipl2)、儿童栏目(For Kids)、青少年栏目(For Teens)。用户可以通过关键词进行检索,检索范围可以进行设置,选项包括 All of ipl2、For Kids、For Teens、Newspapers & Magazines。IPL2 支持“*、?”等通配符(Wildcard Characters)的检索应用,支持“AND、OR、NOT”等布尔逻辑运算符(Boolean Operators),提供范围检索(Range Searches)邻近检索(Proximity Searches)和模糊检索(Fuzzy Searching)。

IPL2 收录的信息资源按照主题被分为艺术与人文、商业与经济、计算机与互联网、教育、娱乐与休闲、法律政府与政治学、科学技术、区域与国家信息、参考资料等 12 个大类,大类下再分小类。

IPL2 延续 LII 采用国际通用的《国会图书馆主题词表》(LCSH)进行标引,但又将 LCSH 中的部分主题改为更符合用户习惯的主题名称,例如,在 LCSH 中用 Electronic Mail System,在 LII 中用 E-mail。它遵守 Z39.50 与 MARC 标准,使得它与其他学科信息门户之间可以实现互操作,这点从 LII 和 ILP 能够比较完美的合并就可以看出。

5) 相关技术的应用

IPL2 注意新技术的应用,采用快速、灵活、功能强大的网页索引系统(Simple Web Indexing System for Humans-Enhanced, 简称为 SWISH-E)。IPL2 已能把用户需要的信息资源和服务有机地集成在一个统一的系统里,并开通了分布式参考咨询系统“Ask an ILP2 Librarian”,为用户向专家交流咨询提供了平台。这些都是新型的学科信息门户应该具备的。提供了社会化网络的分享工具按钮,用户可以将网站分享到 Facebook、Myspace、Twitter、You Tube、Delicious、Diigo、Wordpress 等社会网络空间中。

6) 个性化与人性化服务

ILP2 在用户信息服务只能提供信息的浏览与检索,忽略了用户定制和个性化的服务。此

外，在 Web2.0 环境下，IPL2 网站没有很好的应用相关的技术和应用，没有提供给用户动态的、开放式的数字化空间，也没有提供用户发表评论以及与同行交流的渠道，这些都应该是新型的学科信息门户应该具备的。

7) 更新与维护

IPL2 的信息资源每周更新，工作人员负责新资源的追加和资源链接的有效性。自 20 世纪 90 年代以来，IPL2 与其前身 LIH 和 IPL 一直持续向网络用户提供信息服务，几乎没有间断过，可见，信息的稳定性与安全问题得到了重视。

9.2.4 Web2.0 环境下信息组织评价

在 Web2.0 环境下，网络信息资源的生产、组织、发布与传播更加依赖于用户的作用。大量的用户为网络提供信息内容，成为信息的生产者，并且信息生产具有去中心化特点。用户产生的微内容是网络信息资源的重要组成部分，这些来自用户的微内容通过一定的方式（如人与人之间的社会关系等）进行汇聚，从而形成 Web2.0 网络信息环境，并产生了网络信息组织方式的变革。

1. Web2.0 环境下信息组织特点

在 Web2.0 环境下，传统的网络信息组织方式，如文件组织方式、数据库组织方式、主题目录组织方式和超媒体组织方式等，同样适用；但同时 Web2.0 信息资源的特性呼唤新的组织方式，现有的大众分类（社会化标注）、信息自组织等从不同角度展示了信息结构，较好地满足了用户需求，为 Web2.0 信息资源的组织提供了较有力的技术支持。

在 Web2.0 环境下，基于 RSS/ATOM/RDF/FOAF 等的 XML 数据不再和网站、网页混粘在一起，它们被独立了出来，并能够实现同步、聚合和迁移，通过对 XML 数据的处理，这些内容能自由组合，能被各种应用程序（不论是 Web 程序还是桌面程序）呈现和处理。

在 Web2.0 环境下，网络信息资源组织呈现以下特点：

1) 信息组织以用户为中心

在 Web2.0 环境下，用户在网络信息资源组织中参与性大大增强，深度参与到信息的生产、发布、传播、修改和使用中，既是信息的生产者，也是信息的获取者。网络信息的组织也是从用户的角度出发，如用户可以应用社会标签（Tag）对信息资源进行自由标注。

2) 信息组织方式的多元化

在 Web2.0 环境下，网络信息资源的分散性更加突出，信息组织方式也更趋多元化。出现了以用户体系组织信息的博客（Blog）组织方式，以知识点体系组织信息的维基（Wiki）组织方式，以消除信息孤岛、促进用户之间信息共享的社会标签（Tag）组织方式，以“人际关系”体系组织信息的社会网络（SNS）组织方式等。这些新的信息组织方式并不是孤立的应用，往往呈现出互相融合、互相交叉的复杂、多元化发展趋势，如博客中也存在的类似于社会网络的“博客圈”和社会标签标注等。这些多元的信息组织方式更好地适应和满足了 Web2.0 环境下用户的信息需求，也更好地促进了网络信息资源的序化、有效获取与利用。

3) 以“微内容”为基础的信息组织

微内容是指信息内容分解成很小的单元（类似数据元、知识元、信息元等），一篇文章、一条评论、图片、标签、认识的人都是微内容。在 Web2.0 环境下，网络信息资源以“微内容”为基础，信息组织以用户为中心，网络信息资源围绕用户进行组织和呈现；用户使用信息定制工具，依据自身的价值观来判断信息是否纳入组织体系，以及在组织体系中所处的位置；用户通过相互之间的交流、约束保证信息组织的有效进行，形成有序的网络信息环境。

这种以“微内容”为基础的网络信息组织方式，有效地促进了信息的获取和利用。

4) 强调“人的关系”在信息组织中的作用

在 Web2.0 环境下，互联网更加重视“人”及其“人与人”之间的关系，将“人际关系”作为网络信息组织的一部分。这种“人际关系”不仅包括个人的行为和社会关系，还包括一个人的行为和社会关系与其他人的行为和社会关系之间的相互影响、制约。在 Web2.0 环境下，“信息”与“人”巧妙地结合起来，将“人的关系”作为信息组织的一种纽带。Web2.0 以单个用户为中心，利用人与人之间的关系，通过一定的信用确保机制，将用户组成一张社会性的可信任的信息网络，并将这种人与人之间的社会关系融入到人们获取信息的过程中，使信息更好地围绕用户进行组织。如朋友网中给用户提供了“圈子”、“好友”等页面，这已经成为 Web2.0 网站的基本特征以及最重要的内容组织方式之一，也是社会关系在信息组织中作用的重要体现。

5) 网络信息的自组织性

网络信息自组织是指网络中的信息资源由于用户与用户之间、用户与网络其他要素之间的交互性、相关性、协同性或默契性而形成特定结构和功能的过程，也就是指信息网络无需外界指令而能自行组织信息，自我走向有序化的过程。Web2.0 环境下，互联网就是一个复杂的自组织的信息系统。Web2.0 通过某个主题把用户联系起来，用户之间通过开放式的信息沟通方式分享、互动，建立一定的人际关系，依据某些共享的规则协作推动互联网从无序变为有序。在 Web2.0 环境下，以自组织为中心，个人与个人之间，个人创造的内容与内容之间以及个人汇聚的群体与群体之间，都是以不同的自组织方式架构起来的。信息的自组织方式让人、群体、网络信息内容与应用等充分“动”起来，力量得到了最大程度的爆发。

2. Web2.0 环境下信息资源组织评价

以上综合论述了 Web2.0 环境下网络信息资源组织的特点。但我们必须认识到 Web2.0 并不是一项具体的技术或者应用，而是一种网络发展模式，是一系列相关技术和网络应用的集合，主要包括博客 (Blog)、RSS、维基 (Wiki)、社会标签 (Tag)、社交网站 (SNS)、P2P、即时信息 (IM)、基于地理信息服务 (LBS) 等，它们都为用户提供了生产、组织、发布、更新和共享信息的开放式的技术平台，同时又在管理信息方面呈现各自不同的特色。本节选取博客、社会标签、维基等作为 Web2.0 的典型形式，对它们信息资源组织进行评析。

1) 博客 (Blog) 信息组织评价

博客是目前互联网上信息资源的主流表达方式之一，主要以发表时间顺序对博文进行排列，通过博客管理者自创的分类体系以及社会标签 (Tag) 聚类等进行信息组织。同时，博客还从用户个体的角度出发对信息资源进行自组织。网络用户可以通过 RSS 聚合工具订制自己感兴趣的博文内容，并追踪更新。博客的作用是快速发布观点、体会、新闻或日常记录等，并通过与网络用户的互动、交流丰富信息资源的内容。从效果上看，微内容、交流和围绕相关主题形成资源结构，提高了博文信息组织的质量和针对性。

(1) 评价标准。

博客在网络中是以网站的形式呈现的，因此，9.2.1 节中关于网站信息资源组织评价的标准在这里同样适用。同时，根据博文信息资源组织的特点，也有相应的评价指标。

① 内容质量。

博文内容的权威性如何？博文作者的真实身份及其在相关领域内的知名度或权威性如何？博文首页是否提供博文作者的个人介绍 (包括研究方向、联系方式等)？博文信息是否为其他博文或权威网站摘引、链接与推荐过？

博文内容客观准确性如何？博文内容是否含有政治、意识形态、宗教、商业或其他倾向？

网站在引用其他信息来源时是否注明出处? 信息内容是否符合客观事实, 语义表述是否清晰, 涉及理论、概念、原则、规律等是否准确无误? 博文中是否存在错别字等情况?

原创性指由博客作者写作的原创性博文在该博客所有博文中所占的比例, 能够反映该博客信息内容的独特性。

② 分类与检索。

博客框架结构是否清晰? 提供的信息组织方式如何, 是否按主题、学科等对博文进行分类? 博文分类是否科学、合理、能够被网络用户接受? 是否向网络用户提供方便的检索博客内容的途径?

③ 技术应用。

博客应用 Web2.0 相关技术情况如何? 博客是否使用社会标签 (Tag) 对信息资源进行标引, 是否制作了标签云? 博客页面是否添加醒目的 RSS 订阅按钮, 是否提供多种的 RSS 订阅方式, 每篇日志最后是否提供 RSS Feed 链接? 是否是通过社会网络工具进行网络用户信息、博客内容信息的聚合?

④ 互动交流。

博客是否提供了多种与网络用户进行信息交流的渠道? 网络用户是否可以把使用过程中的问题和疑惑告知信息资源的管理者, 信息资源的管理者是否可以及时解决这些问题, 回答使用者的疑惑? 博客是否能够提供社会网络空间? 是否提供社会网络分享工具?

⑤ 更新频率。

更新频率直接体现网络信息资源的时效性, 是评价博客的重要指标, 一个博客更新信息的频率越快, 它所提供的信息的时效性就越强, 利用价值就越大。博客内容是否得到持续、及时的更新? 更新周期有多长? 一般来说, 运行状态稳定的博客应符合至少每周更新一次的要求。

⑥ 稳定性。

由于博客运行商或博客作者自身的诸多原因, 博客经常会出现“搬家”、关闭等现象。

“搬家”是指博客作者将博客内容从一家博客平台迁移到另一个博客平台, 在迁移的过程中, 能否完整地保留博文和评论是评价一个博客稳定性的重要标准。博客的运行状态是否稳定? 博客站点是否稳定? 博文是否进行了归类和存档? 当出现博客站点更换、博客内容搬家等情况时, 在原博客首页是否有明显的告知信息, 并提供新站点的链接?

(2) IFLA 官方博客评析。

IFLA Blogs (<http://blogs.ifla.org/>) 是国际图联 (International Federation of Library Associations and Institutions, 简称 IFLA) 开设的官方博客, 采用了开源的 WordPress 系统。WordPress 是一种使用 PHP 语言开发的博客平台, 用户可以在支持 PHP 和 MySQL 数据库的服务器上架设自己的博客。

IFLA 的诸多分部和小组都开设了博客, 并集成在 IFLA Blogs 中。目前, IFLA Blogs 一共拥有 27 个分部和小组的子博客, 如采访与馆藏发展分部博客 (ACD Blog, by IFLA Acquisition & Collection Development Section)、学术与研究图书馆分部博客 (Academic and Research Libraries Section Blog)、公共图书馆分部博客 (IFLA Public Libraries Section Blog)、参考咨询与信息服务博客 (IFLA Reference and Information Services Blog), 等等。在 IFLA Blogs 首页上, 列出了 27 个子博客的链接、最新的 15 条博文 (信息描述包括博文的题目, 发布时间, 来源子博客名称)、信息搜索 (可以对 27 个子博客的内容进行关键词检索) 以及最近博文的 RSS Feed 等。下面根据以上提出的博客评价指标对“IFLA 采访与馆藏发展分部博客” (<http://blogs.ifla.org/acd/>) 进行评析。

① 内容质量。

IFLA 官方博客采用统一页面设计,并且整体风格、Logo 使用和颜色选择等 VI (Visual Identity) 视觉设计都与 IFLA 官方网站保持一致,博客网站页面结构简单明了、字体大小适中;博客页面上没有放置广告,符合作为专业性学术博客的特点。

IFLA 是世界图书馆界最具权威、最有影响力的非政府专业性国际组织,作为其官方博客,所发布的信息在权威性和准确性方面是毋庸置疑的,可以成为用户获取相关领域最新信息的重要渠道。IFLA 采访与馆藏发展分部博客主要发布该分部工作与项目开展的最新动态以及一些观点和态度,成为与图书馆员和信息专家交流思想的平台,该博客主要发布采访与馆藏建设相关的信息,包括通告、会议信息、课程信息、出版物、论文等。

笔者浏览了 IFLA 采访与馆藏发展分部博客自 2012 年 5 月 6 日开设至 2013 年 8 月 31 日发布的 247 篇博文,发现论文、会议信息与课程信息等收集转发的信息资源占了将近 80%,可见该博客信息资源的原创性不高。但同时我们也要看到,这些收集自网络的博文都是经过博客管理者精选和分类,并围绕图书馆采访与馆藏建设的主题进行了类聚,对于用户了解和掌握该领域的最新动态具有很高的价值。

② 分类与检索。

IFLA 采访与馆藏发展分部博客的分类体系的目录包括:ACD 通告 (ACD Announcements)、ACD 出版物 (ACD Publications)、ACD 评论 (ACD Reviews)、征文信息 (Calls for Papers)、会议论文 (Conferences Proceedings)、课程信息 (Courses & Online Courses)、综合 (General)、一般论文 (General Articles)、即将举行的会议 (Upcoming Conferences)。分类体系主要是按照信息资源的类型设计的,没有进行学科主题的分类,这对网络用户通过分类体系浏览博文信息存在一定的障碍。

该博客除了对博文进行分类外,还标注了社会标签 (Tags),每条博客所拥有的标签数量不一。虽然博客中注明的是“主题词” (Topics),实则是社会标签,原因在于博客发布者所标注的词一般都是来源于自然语言,较为随意,并不严格限于某个受控词表。如点击标签“Big Data”,就可以显示所有被标注了 Big Data 的博文。

该博客提供了“按月归档”,实际是一种通过时间顺序,以月为单位对博文进行组织的途径,用户也可以以此浏览博客。

该博客的搜索功能支持关键词模糊检索,检索结果显示命中条数、博文标题、发表者、发布日期、类别、主题以及部分博文内容等,可以通过点击“Read full post...”进行全文阅读,每页显示 7 条记录,用户不能选择检索结果显示条数;也没有提供高级检索。

提供了“条目 RSS”、“评论 RSS”,博客已将博文和评论发布成 XML 格式,方便用户订阅,也方便数据被其他信息系统使用,具有较好的互操作性。

③ 技术应用。

IFLA 采访与馆藏发展分部博客应用了社会标签技术,博客作者对每条博文进行了标注,并在网站上形成了标签云,常用标签按照首字母排列顺序,并通过标签字体大小可以判断其出现频率的高低,网络用户可以通过标签云检索博文;提供博文、评论的 RSS 聚合,方便网络用户的订阅;应用了为搜索引擎优化而提供的永久链接 (PermaLink) 系统;提供对博文进行嵌套分类的功能,同一博文可属于多个分类,如“CILIP's Umbrella Conference 2013”既被归于综合 (General) 类,又被划分在即将举行的会议 (Upcoming Conferences) 类。

④ 互动交流。

IFLA 采访与馆藏发展分部博客提供了用户评论接口,用户可以围绕相关博文在线发表评论,与博客作者进行交流。据笔者的统计,自从博客发布信息以来,用户在博文后留言只有两条,并且很多博文页面的评论窗口已经被关闭。由此可见,该博客的交流功能还有待加强。

该博客没有利用社会网络应用和工具加强其与用户之间以及用户与用户之间的交流互动,也没有提供博客信息资源的分享机制和途径,用户不方便对博客信息资源进行转发、分享等操作,不利于信息资源的共享。

⑤ 更新频率。

截止 2013 年 8 月 31 日,IFLA 采访与馆藏发展分部博客一共发表博文 247 篇,创建以来平均两天发表一篇博文,更新频率较高。对于用户而言,可以经常从该博客中获取新的专业性信息资源。

⑥ 稳定性。

IFLA Blogs 是 IFLA 官方网站的二级网站,其运行的稳定性能够得到充分的保障。

2) 标签的信息组织的评价

标签(Tag)是 Web2.0 网站广泛使用的信息资源组织方式。Tag 标引是一种自由而有序的信息资源分类技术,是网络用户运用自定义 Tag 的方式进行协作分类的活动。基于 Tag 的信息组织方式允许网络用户根据自己的兴趣和需要用个性化的语言自由创建 Tag、自由标引各类信息资源(如文档、图片、视频等文件),对信息资源进行分类管理,实现有效的信息检索和社会化传播。不同用户使用相同的 Tag 描述、组织相关内容的信息资源,则可以将这些信息资源聚合起来,促进了信息资源的共享,避免“信息孤岛”的出现。

(1) 基于标签的信息组织特点。

① 存在个性化与有效共享的矛盾。

在网络信息资源的组织实践中,基于标签的信息组织方式存在个性化与有效共享之间的矛盾。

一方面,个性化是标签的主要优势之一,网络用户完全可以从自身的角度出发,利用 Tag 实现对信息资源的揭示、标引和组织,充分满足用户个性化揭示与组织信息资源的需求,有利于用户对信息资源的利用。

另一方面,因用户在创建与使用标签过程中过于随意和个性化,不同的用户或同一用户的不同阶段对于同一信息资源具有完全不同的理解和完全不同的表达,造成阶段标注同一资源使用的标签差异性较大,难以形成信息资源的有效共享。

因此,标签信息组织机制及其相关处理技术需要着重解决标签使用过程中个性化与有效共享之间的矛盾,既能保证用户使用标签标引的个性化,同时不失去共享性,实现两者之间的平衡。

② 信息结构的平面化。

基于标签的信息资源组织,词间没有相关关系和等级层次,从结构上说是平面的。这种平面结构关系简单明晰,便于网络用户的平面浏览。标签之间缺乏聚类成族能力,同义和近义标签无法组织或联系在一起,因此,标签不能作为大型网站信息资源集成组织的主要途径,也不适合小型网站长期积累大量信息资源的组织,只适应小型网站的临时性资源组织或者网站辅助性的信息资源组织途径。

标签云(Tag Cloud)是一个结构松散的平面化的信息资源组织工具,本质上是一个平面化的资源浏览窗口,可容纳的标签数量有限,它难以适应信息资源结构复杂的大型网站的信息组织。目前,大多数网站采用栏目分类或主题树加局部标签云的方式弥补标签组织的平面化缺点,大多数博客也主要是以栏目分类组织信息资源,这说明了标签云简单平面化结构特征限制它的广泛使用。从资信息源集成和标签聚类来看,长期积聚的信息资源通过个性化松散标签来表达是不可能的,因此,对大量标签实行集成聚类,形成等级式的主题树是主要的解决办法。

(2) 评价标准。

在 Web2.0 环境下, 标签不是以独立的形态在网络中呈现, 而是依附于门户网站、博客等网络载体, 作为信息组织的工具。因此, 我们对标签信息组织进行评价, 主要是针对内容状况和实现功能展开。

① 标签的内容状况。

用户标注的标签是否存在错误? 由于标签标注是网络用户自主行为, 受用户文化程度、认知状况等的影响, 信息资源的标签中会出现错别字的情况。

标签的重复度如何? 所谓标签的重复并非指两个标签完全一样, 可能是中文简体字与繁体字的重复, 也可能是不同语种的语义重复。

标签的专指度如何? 标签的准确性如何? 标签是否能够反映被标注对象的主题特征, 能否帮助其他用户获取与其需求相匹配的信息资源?

② 标签系统的功能。

标签系统是否提供对词汇的控制? 标签主要来自于自然语言, 存在着同义和异义等状况, 使得信息资源按照标签的聚合会出现偏差, 导致检全率和检准率不高。标签系统可以在后台构建一个对标签进行后控的词汇库, 其内容包括基本的同义词词表和多义词词表。

标签系统是否构建了标签的层次体系结构? 结构的平面化导致标签缺乏上下级之间的类别管理, 标签之间的关联性弱。标签系统可以对标签进行分类, 标签与标签之间就不再是一个平面结构, 而是有了比较明晰的等级层次关系。

标签系统是否提供标签的质量控制? 由于用户对信息资源标签标注行为自由度较大, 这样会导致标签的良莠不齐, 需要标签系统提供质量控制的机制。

是否构建标签标注用户的社群, 实现用户之间的交流与共享? 标签系统的社区聚合功能能够很好地将用户与用户、用户与资源用标签联系起来, 构成了一定规模的社会网络, 并且利用标注系统的运作原理不断强化和扩展已经形成的社会网络。

标签系统是否提供与其他标注系统的互操作? 是否提供标注系统的 API 接口?

(3) 豆瓣网标签评析。

豆瓣网是一个集博客、交友、小组、收藏于一体的新型社区网络, 已被公认为中国 Web2.0 时代最纯粹最精彩的先锋网站。作为大众标注网站的一个典型代表, 豆瓣网的标注范围主要包括图书、音乐、影视和博客等领域。

① 标签的内容状况。

在豆瓣小组中, 小组的标签完全由管理员自行决定, 而管理员因受其文化程度、知识水平以及其他方面情况的限制, 很容易出现漏标、误标、滥标或赋予的标签太过于个性化而通用性较差等问题, 从而导致其他网络用户通过标签进行检索时不一定能获取准确的信息资源。

在豆瓣读书、豆瓣电影和豆瓣音乐中, 存在标签冗余、语义重复、专指度不高、缺乏准确性等问题。如《张爱玲文集》同时具有“文学”和“Literature”两个标签, 而这两个不同语种的标签在语义上就有重复;《不能承受的生命之轻》的常用标签中出现了“米兰·昆德拉”、“米兰昆德拉”和“昆德拉”, 从含义上说, 这三个标签是重复的, 从书写规范来看, 后两个标签明显不符合外国人名的书写规范; 武侠小说《射雕英雄传》用户常用的标签中同时含有“武侠”、“小说”和“武侠小说”3个标签, 这3个标签在语义上有交集, 存在包含关系, 而导致标签的专指度不高; 电影《致我们终将逝去的青春》用户常用标签中有一个是“2013”, 这个标签缺乏准确性, 并且专指度不高, 不能反映该电影的主题特征。不规范的标签会占用大量的资源, 一方面给标签的管理带来了诸多不便, 另一方面会造成网络用户在标签海洋中的迷失。

在 Web2.0 的环境下, 豆瓣网中标签的规范化既是当前的热点问题, 也是难点问题, 急需加以研究解决。

② 标签系统的功能。

豆瓣网标签系统没有建立后控词汇库, 不能对标注标签词汇进行控制, 用户可以自由地选择标签。

豆瓣网对用户标注频率较高的标签进行了分类, 包括文学、流行、文化、生活、经管、科技 6 大类, 大类下涵盖 145 个热门标签。虽然这种分类较为简单, 但在一定程度上改变了标签结构平面化的缺陷, 有利于用户通过标签查找信息资源。

在豆瓣网中, 当用户对一资源对象进行标签标注时, 系统会提供该资源对象的 10 个“常用标签”供用户参考; 此外, 具体的资源页面上, 按照被用户标注的次数从多到少列举了 8 个“常用标签”。一般来说, 如果一个标签被用户广泛地标注, 就是“群体智慧”的反映, 具有代表性。这种“常用标签”的推荐可以看作是对标签质量进行自组织控制的一种方式。

豆瓣网本身就是一个社会网络, 用户之间可以进行充分信息资源交流与共享。用户完成对一资源对象进行标签标注后, 系统可以将相关信息分享到“豆瓣广播”, 也可以通过去绑定新浪、腾讯等微博, 将信息分享到其他的社会网络空间。

豆瓣网现在已经开通了应用程序接口 (Application Programming Interface, 简称 API) 服务, 为第三方开发人员提供编程接口。网络用户可以利用豆瓣的数据和功能, 不过目前还没有提供专门针对标签的 API 功能。

3) 维基信息组织的评价

维基 (Wiki) 是基于网络用户集体智慧形成的人类知识网络系统。用户之间通过开放式的互动协作共同生产、组织维基信息资源, 以知识点体系组织信息, 针对同一主题作外延式和内涵式的扩展, 实现知识的有效组织、管理和利用, 进而实现知识的共享与创新, 并能通过主题将用户联系起来。维基信息组织方式充分体现了用户个体之间的协作在互联网信息资源组织中的重要作用。

(1) 维基信息组织特点。

① 开放协作。

维基具有开放性, 提供了“共同创作” (Collaborative) 的环境, 赋予用户相当大的编辑修改权, 可以在 Web 的基础上对维基文本进行浏览、创建、修改。维基可以帮助用户分享和利用社群内某个领域的知识, 用户对维基内容有很大的权限, 可以进行阅读、下载、创建和修改, 自由开放度很高。

维基一般坚持开放多元的中立观点, 编辑方针采用中性观点 NPOV (Neutral Point Of View), 要求必须能包容所有人的观点, 以实现其包容性。在实际操作中, 维基把不同的观点和事实以及它们的变化都记录下来, 而不判断“正确”和“错误”, 或根据这种判断隐藏“错误”的观点和事实。开放多元的定位为其实现自组织管理和自组织发展演进奠定了基础。

② 信息自组织。

信息自组织是维基重要的特征。与传统模式下通过他组织实现信息有序化不同, 维基是通过用户的协作共享, 以自组织方式实现信息有序化的, 这种信息的共享协作是维基能够不断演进的动力。因此, 维基的有组织状态是一种动态平衡状态, 向着有序化程度更高的水平和层级演化。不但维基页面的内容不断改进, 水平不断提高, 整个超文本的组织结构也是可以修改、演化的。同时其内容也会逐渐形成会聚, 系统内多个内容重复的页面可以被汇聚于其中的某个页面, 相应的链接结构也随之改变。

(2) 评价标准。

维基本身是一个网站, 因此 9.2.1 节中关于网站信息资源组织评价的标准在这里同样适

用。同时, 维基作为 Web2.0 的典型代表形式, 根据其自身的特点也相应应有信息资源组织评价指标。

① 资源规模与内容。

维基的信息资源规模如何? 容纳了多少条目, 创建了多少内容页面? 涵盖的学科领域如何? 维基条目内容描述是否详细? 维基是一种网络百科全书, 丰富和全面的信息资源内容是满足网络用户信息需求的基本前提, 也是进行信息资源组织的基础。

② 资源组织与检索。

维基网站是否提供多元化的信息组织方式, 是否具有完善的检索途径? 维基信息组织层级是否合理? 信息组织层级是指网站组织具体信息的结构层次, 简单地讲就是用户进入一个网站的主页或者网页后, 需要点击几次鼠标才能找到具体的信息。信息组织层级越多, 信息资源被细分化的程度就越高, 但用户检索起来比较麻烦。一般来说, 2~3 层的信息组织层级比较适合用户的使用。

③ 用户参与。

维基网站是否通过多种途径为用户参与信息资源建设提供便利? 维基网站有多少用户注册? 活跃用户的多少? 维基被用户编辑了多少次? 每一页平均被编辑次数是多少? Web2.0 环境下信息组织的显著特点就是用户的广泛参与。维基作为 Web2.0 的典型代表, 用户的参与度如何, 是评价维基网站的重要指标。

④ 质量控制。

维基作为一种网络百科全书, 需为用户提供内容正确、准确与完整的信息资源。维基网站是否提供信息资源质量控制机制? 通过何种途径和方法确保信息资源的内容质量? 对于用户参与维基条目的创建、编辑等是否进行权限分级管理? 维基网站是否有相关的技术和规范保证信息资源的有序化和用户行为的规范化? 对于出现一词多义的情况, 如何进行歧义的消除?

⑤ 共享技术应用。

维基网站是否应用相关技术为信息资源的共享提供便利? 是否提供 RSS 的订阅? 是否开放信息资源的 API 接口? 是否开设了维基网站的移动版本, 是否开发了维基的相关网络应用和手机客户端?

(3) 中文维基百科评析。

中文维基百科是中文领域具有代表性的集体创造知识的协作平台之一, 它是维基百科协作计划的中文版本。下面根据以上提出的维基评价指标对中文维基百科进行评析。

① 资源规模与内容。

中文维基百科是不断发展、不断增长的有机体。用户可以不断创建新的页面, 从而使系统得到增长。2002 年 10 月 24 日, 中文维基百科第一个条目发布, 10 余年来得到迅速扩展。根据中文维基百科首页的统计数据, 截至 2013 年 8 月 31 日, 一共有 722 398 篇条目, 页面数量达到 3 070 945 个, 内容主题也覆盖了广泛的学科领域。但是, 相对于百度百科的 626 万余个条目而言, 中文维基百科条目数量的增长较慢。

中文维基百科条目内容描述较为详尽, 包含概述、详述、目录、参考资料、扩展条目集、分类、统计、编辑等内容, 另外还罗列了注释、用户评分栏目, 但不同页面之间设置差异较大。中文维基百科还包含“阅读”、“编辑”、“查看历史”、“讨论”页面, 其中“阅读”页面包含条目的概述、详述、目录、参考资料(脚注、参考文献、参见、外部链接、多媒体资源)、扩展条目集、分类、注释和条目评分; “编辑”页面供所有用户编辑(包括匿名用户), “查看历史”页面主要包括历史统计(如浏览次数、编辑次数等)、历史版本、贡献者, 还可以进行历史版本的比较; “讨论”页面, 主要记录用户对该条目的讨论话题, 这个页面也有自

己的子页面“阅读”、“编辑”、“添加新话题”、“查看历史”。同时,中文维基百科提供大陆简体、港澳繁体、新马简体、台湾正体4个版本供选择,以扩大其在全球范围的影响力。

② 资源组织与检索。

中文维基百科提供了多种条目组织方式,以帮助用户从不同入口找到所需条目,包括页面分类、分类索引、常用列表和主题索引四类。

页面分类对条目进行多角度归类,当用户通过分类索引浏览某一条目时,系统会指示该条目页面分类的位置,用户可以随意链接到该条目对应的各级上位类。

分类索引是在页面分类的基础上形成的用于用户分类浏览的分类体系。分类索引有8个大类,二级14个、三级类目3088个。其中“自然与自然科学”类目最大,其次是“社会”、“生活、艺术与文化”类目。中文维基百科的类目较深,对条目分类的较为仔细,深度为10级左右,一个最低层类目下的平均条目为20条左右,这样导致1199多个类目对应庞大的条目数6261743个(截止2013年8月31日),平均一个类目下有5200余个条目。从一级类目看,中文维基百科则采用突出列类的方式,将重要的类目单独设类,如“中华文化”、“世界各地”类目分别与“生活、艺术与文化”、“人文与社会科学”类重合。由于百度百科的开放分类最多只有三级类目,且中文维基百科从四级类目开始,类目名称中用自然语言的比例大大增加,规范程度较低。

常用列表将各大类目中可以用年表、列表方式组织的类目集中排列,包括历史年表、化学品列表、世界体育竞赛列表、国际组织列表、遗产列表、世界地理索引、语言列表等常用列表,列表参考杜威十进制图书分类法进行了分类,用户通过浏览可获取自己感兴趣的词条。

主题索引是按照用户感兴趣的某一类主题组织页面,并围绕该主题将所有相关的条目罗列,供用户阅读及做进一步深入链接,目前中文维基百科提供了200个主题导航,通过与分类索引的类目进行比较,发现主题索引的12个大类中有7个大类与分类索引基本相同。如“艺术与文化”、“数学与科学”、“技术”、“社会科学”、“人文”、“宗教”、“地理”。另外5个大类选择日常生活中人们感兴趣的主题单列出来作为大类,包括:“娱乐与体育”、“气象”、“交通”、“军事及战争”、“历史”。其中很多都是分类索引的二级类。这种以事物为核心的组织方法,弥补了分类索引学科分类法的不足。

用户可以选择以上四种途径,通过浏览的方式找到所需的条目,也可以通过输入关键词进行检索获取相关条目。关键字检索与等级目录查找是中文维基百科相互补充的两种条目检索形式。但是,在这两种检索形式均存在一定的不足,如:分类目录层次太多,不利于用户查找相关条目;检索系统的智能化程度不高,关键词检索不支持较长的提问式查询,增加了用户在检索过程中的复杂度。

③ 用户参与。

中文维基百科为了实现信息的开放共享,吸引更多用户参与,采用简单的“格式化语法(Formatting Rule)”取代HTML的复杂格式标记,类似所见即所得的风格,用户可以快速创建、存取、更改超文本页面,而且链接方便,通过简单标记,可直接以关键字名来建立各种链接(包括页面、外部连接、图像等);同时,系统还支持面向社群的协作式写作,并为写作提供帮助。

截至2013年8月31日,中文维基百科的注册用户数为1488675,自维基百科建立以来页面编辑次数达到29429054,每页平均编辑数9.58,活跃用户7516个(在前30天中操作过的用户)。相比而言,百度百科的注册用户数已经达到了340余万,

④ 质量控制。

中文维基百科的用户拥有对条目进行编辑、分类、评价等权利,但是对于信息资源影响力较大的操作(如创建、删除条目)则需要由系统管理员进行审核。此外,系统管理员对

用户拟定了层次放权的方式,不同级别的用户对条目的处理有不同权限;系统管理员通过设置简单的格式化语法让用户对其上传的信息进行规范化;在对信息的整合上,中文维基百科持中立观点对不同用户上传的不同信息采用保留的措施,保存页面每一次更动的版本并进行展示。

中文维基百科注册用户可以任意创建、修改、删除页面。在修正错误方面,中文维基百科采取了一种与权威校正不同新的质量保障方式,即力图做到让“更多的眼睛发现更多的错误”,一篇文章往往由上百人审查,而且是在不断讨论的基础上接受检阅,只要有足够多的读者关注,所有的条目问题最终都会被发现并得到纠正,大家最终会将一个文档修改得更好,而不是更差。中文维基百科拥有“修订版控制系统”(Revision Control System,简称 RCS)管理条目内容,用户可以随时找回以前的数据并对比。维基没有最终的正确条目,它永远都是在不断地丰富和完善,参与的人越多就会越完善。

中文维基百科通过用户互动形成自组织系统规范,自组织的过程也是对信息资源质量进行控制的过程。通过用户的参与,并经过用户协同的检验,维基百科形成了成熟的技术和规范,包括:保留页面每一次更动的版本;可以进行版本对比,每个页面都有更新纪录;用户更新页面时可以在描述栏中说明,使管理员知道更新页面的情况;初次参与的用户可先使用沙盒(Sand Box)页面做测试;具有有记录和封存 IP 的功能,使破坏者、恶作剧者无法胡作非为;极端情形下可以用锁定技术将内容锁定,其他用户无法再编辑;通过用户的互动,维基百科形成了相对稳定的编辑规则,这些规则一旦形成就对使用者具有约束力。

中文维基百科通过建立消歧义页来实现消除由于一词多义所引起的歧义。如大学既是教育机构又是一本古书,博士既是官名,又是博学多才的人,又是师傅,还是一种学位的名称,在这种情况下就有必要进行歧义的消除。消歧义页中罗列了有相同或相近的标题,但内容不同的条目,如“奥巴马”的消歧义页列举了“老巴拉克·奥巴马”、“巴拉克·奥巴马”和“米歇尔·奥巴马”三个条目,并进行了区分。

由于用户可以自由地对中文维基百科条目进行修改,并且内容不一定会受到严格的审核,或者受到系统管理员认知程度的限制,因此中文维基百科不能保证每条条目都符合某些社会规范、内容精确无误。

⑤ 共享技术应用。

维基百科提供了 RSS 推送订阅服务,用户可通过在维基百科订阅 RSS 或者 Atom 的方式,无需在网页中打开中文维基百科就可以及时跟踪页面的修改以及内容的更新。中文维基百科开设官方 WAP(无线应用协议)版网站,以供用户在手机及个人数码助理上阅读和使用维基百科。另外,维基百科还开发了基于 Android、iOS、Windows 8 和 Firefox OS 的维基百科应用程序(Wikipedia App),方便用户在不同的移动设备操作系统上,都能随时浏览维基百科。相比而言,百度百科则进行了更多的技术应用,促进信息资源的共享,如:百度百科的 API 输出,合作用户可在网站或程序中使用百度百科的数据资源,为合作方用户提供更全面、客观、及时的信息服务;基于 Android 和 iOS 开发了百度百科手机客户端;建立了百度百科的 WAP 版;应用了二维码技术,方便移动设备获取百度百科的信息资源;可以通过“百度分享”将词条信息分享到 QQ 空间、新浪微博、豆瓣网、Facebook、Twitter、Linkdein 等用户常用的社会网络空间中。

9.2.5 数字图书馆信息资源组织的实例与评价

随着计算机技术和网络技术的发展,网络的互联使访问分散在各处的信息资源成为可能,为了提高网络资源的序化程度和便于人们的利用,文件、超媒体、数据库、网站、网络

资源指南、搜索引擎、编目和学科信息门户信息组织方式已运用于 Internet 信息的组织。从某个局部来看,如某个文件、利用超文本技术相链接的相关资源、某个网站、某个数据库、某个编目记录集合或某个学科信息门户,是有控制的、相对集中的、有序和规范的。但从总体上看,由于互联网上的信息没有统一的控制,信息的质量参差不齐,网上的信息是分散、无序、不规范的;由网络互连在一起的分布信息仓储是异构的,这些各自独立的信息仓储具有各自不同的组织、描述和检索方式,难以实现跨仓储的统一利用;对知识的运用还远远不够,尤其是面向用户需求的知识和领域知识。人们需要一种跨仓储的、统一的、高效的访问和利用工具,以及高质量信息的生成、组织和提取途径,数字图书馆恰恰迎合了这种需要。

1. 数字图书馆及其信息组织的特点

目前,国内外还没有形成公认的数字图书馆的定义。

笔者认为,数字图书馆是以数字化资源为馆藏,以先进的信息处理技术与计算机设备为手段,以国际互联网为服务平台,以信息收集、开发、管理、存储并提供利用为己任的分布式、面向对象的巨型数字信息空间。作为最为复杂和系统的网络信息资源组织方式,数字图书馆信息组织的特点如下。

1) 复杂的结构

信息组织结构在这里是指在数字图书馆中组织信息的结构,其设计直接影响着数字图书馆中数字资源的存储、管理和检索方式,是数字图书馆体系结构中的一个核心部分。在构建数字图书馆信息结构时须考虑三个简单概念。① 数据类型:描述数据的技术属性,如格式和处理方法。② 结构元数据:描述数字资料的类型、版本、关系等特性的元数据。③ 元对象:提供对数字对象集的引用。最简单的元对象是一个指向其他数字对象的句柄的列表。如列出某物理项的所有数字化版本的数字对象即是一个元对象,被用来聚集数字对象,如用元对象来区分两个不同版本的扫描图片。

数字图书馆信息结构的构建包含两个含义:一是数字对象的组织结构,数字对象是数字仓储中表示信息的基本逻辑单位,如一篇文章、一张图片、一部音乐作品或一段影像,数字对象的信息结构决定着进一步的信息组织、处理和利用方式;二是分布式信息仓储的组织结构。数字图书馆的收藏可以特指本地的信息仓储,也可以是互连的信息仓储的集合。如何建立一个统一的、互操作的、可伸缩的组织框架,将分布互连的信息仓储集成为一个整体,在此基础上提供高质量的信息服务,如屏蔽各仓储的差异,提供统一的服务接口、语义化检索、智能代理等,是确定分布式信息仓储的组织结构时要解决的问题。

构建数字图书馆信息结构的要求是:提供有效组织信息的途径,其组织方式为计算机所理解并完成用户所希望的交互,具有提供用户及其应用以相当的灵活性,使收藏的数字资源易于管理,及时反映信息基础结构在经济、社会和法律等方面的发展。

2) 海量的信息

时至今日,信息存储的度量单位已经由 KB、MB 发展到 GB、TB 甚至 PB。数字图书馆面临的数据是多类型和海量的,它存储的信息是以 TB (Terabyte, 1TB=1024GB) 为最低限度的,如“美国的记忆”已超过 104TB 的存储量。中国数字图书馆工程使用并行数据库技术和分布式计算机系统来支撑海量的元数据系统。

数字图书馆的信息种类繁多,数量庞大,必须采用自动化的信息组织手段,借助海量信息存储技术。按照加拿大北电网络公司已经开发出的互联网用高速激光器计算,借助光纤每秒传送 400 亿位数据,将全球最大的美国国会图书馆馆藏文献传送到北美各地,只需要 3 秒的时间。根据这一对信息存储与传输能力的技术预测,数字图书馆存储海量信息的能力将不断增强。

3) 分布的资源

数字图书馆资源分布化体现在：它们存放在不同结构的不同空间中，由于数字图书馆是一个开放式的系统，构成数字图书馆数据层的各个“存储小间”(Storage Cells)有着不同的目标和存储对象，每个仓储在本地对各自的信息进行组织，并施以相应的筛选、索引、联合等控制，由不同的单位(机构)建设或管理，在此基础上，再在各信息仓储间进行互连。在总体上构成一个分布式的系统。该系统要涵盖多个分布式的、超大规模的、具有可互操作的异构多媒体资源库群，通过互联网对国内外用户提供高效、跨库、无缝链接的信息服务。

从管理的角度看，分布式管理是数字图书馆发展的高级阶段，它意味着全球数字图书馆遵循统一的访问协议之后，数字图书馆可以实现“联邦检索”(Federated Search)，全球数字图书馆将像现在的 Internet 一样，把全球的数字化资源连为一体，连接成为一个巨大的图书馆。数字图书馆将成为未来社会的公共信息中心和枢纽，实现多媒体存取、远程网络传输、智能化检索、跨库无缝链接、创造出超时空信息服务的新境界，数字图书馆的服务将延伸到每个人的桌面。

2. 数字图书馆信息组织的评价指标

数字图书馆的上述特征说明，数字图书馆的信息组织必然有别于前文所讨论的各种组织方法，其评价标准和要求也更高。

1) 信息组织对象的采集

相对于其他的组织方式，数字图书馆的信息采集渠道更多。从国际上的实例看，数字图书馆的对象来源已远不止将现有的已出版的资料数字化，而是扩大到包括数据集、网页、教学材料、试验数据、博物馆物品、出版社的原始材料、对专家用户追踪所得的信息等。

是否收集现有的数字化信息？包括图书馆已建立的电子化、网络化的书目数据库与全文数据库系统，出版商出版的电子出版物等。

是否将网上的信息下载到本地数据库？网上的数字化信息越来越多，如网上电子书、电子期刊、电子报纸、网站等，数字化图书馆信息组织的主要内容就是收集互联网上或其他来源的数字化信息。这种采集方法经济、省时省力。是否鼓励用户推荐和补充资源？下载的资源是否侵犯了知识产权？这些信息是否来自固定的信息源？(固定信息源是指那些能长期提供信息、信息的内容范围较为稳定并能经常进行维护、且不时增加新信息的网站，如政府机构、学术团体、图书馆、出版社、网络公司、数据公司等。)对数字图书馆来说，应该选择固定信息源，优先选择图书馆、出版社、政府机构及有一定历史和信誉的机构(如 OCLC)与网络公司(如 Google, Yahoo!)，同时要时刻关注和研究网站的变化情况，随时增加新的固定信息源。

是否对馆藏特色文献进行数字化处理？即借助专用设备(扫描仪、数码相机等)与软件将非数字化信息转换成数字化信息，用标记语言编辑上网，如上海图书馆将大量古籍文献转换为数字化信息。转换过程需要投入巨大的人力物力，但对于“激活”馆藏珍贵文献的保存与开发利用是功德无量的工作。这种采集方法获得的资源具有可供全文检索、节省空间、便于网络传送等优点，但制作成本高，速度慢，需要花费大量的时间和精力进行校对。

2) 信息的揭示与描述

数字图书馆是否将采集到的、原先存在于不同类型信息载体的信息对象组织在同一个界面？在数字图书馆中，信息的类型是多种多样的，同时这些信息的来源也很广，有的在数据库里，有的则存放在光盘上，有的存放在磁带机上，数字图书馆要撇开信息的类型差异，统一将它们作为数字对象进行组织。

能否对不同媒体、不同格式的信息对象进行揭示？除了文字信息外，声音、图像、视频

等类型的信息,只要能够数字化,也都可以是数字图书馆的收藏和组织对象。如上海数字图书馆项目组织丰富的馆藏文献,包括古籍、民国图书、地方文献、科技报告、中外期刊、音响资料、历史照片等数以万计的各种资料,按照读者需求和文献特征形成9大系列,即上海图典、上海文典、点曲台、古籍善本、科技会议录、中国报刊、民国图书、西文期刊目次、科技百花园。

是否能够对各种来源渠道的资源进行整合?数字化图书馆不仅是数字化信息的组织者,同时也要将数字化信息与非数字化信息进行有机的集成,组成一个广泛的、有序的和完整的信息组织体系。由于传统书目信息和数字化图书馆的信息组织在技术和标准等方面有较大的差别,能否建立起这两种信息组织之间的对应关系,是数字化图书馆信息组织中值得注意的问题。

对信息的揭示是否是多维的?多角度地揭示数字化信息是数字图书馆提高信息检索与利用效率的基础,多维的信息揭示便于从多种角度来描述信息的特征,如文字信息可以从作者、题目、出版商等多种角度来反映该信息的特征,而声音信息可以从音色、音质、音量等方面揭示。过去,人们习惯于利用分类语言和主题描述语言对各类信息对象的内部特征和外部特征进行描述,但这样的揭示只局限于题名、关键词、分类号等有限的几个方面,而都柏林核心元数据(Dublin Core)15个基本元素的利用使得对数字信息的揭示可以从15个方面进行,而且这些元素还可以重复。

是否对信息进行非线性组织?信息的非线性组织将信息组织成一个网状结构,该信息网中的任何一个信息单元都由一组与其相关联的信息点连接着。对任何一个信息单元的检索都可带动其他若干信息单元的搜寻,读者可以在各个信息单元中自由切换。

是否遵循标准化原则?包括数据格式的标准化、描述语言的标准化和标引语言的标准化。如果没有一个统一的分类标准和索引方法,将来开发数字图书馆信息的检索工具就会非常困难,势必出现针对不同的揭示方法制作不同的检索工具的混乱局面。另外,数字图书馆的数据类型也不同,如文本信息、地图信息、图像信息及视频、音频等信息,对不同的内容,需要不同的分类体系和索引机制。而能否制定一个比较好的分类方法、建立一个比较好的索引机制,将直接影响到后续工序。

在数据著录格式方面,其数据库是否不仅支持已有的国际标准(如ISO 2709, MARC)和国内标准(CNMARC),而且支持最新的RDF和XML格式?

是否利用各种自动化处理技术?包括自动标引、自动文摘生成、自动篇名生成、自动分类等。由于数字图书馆信息规模庞大,完全依赖手工分类与标引是不切实际的,所以尤其需要在自动化标引与组织方面进行更多的研究。目前,国际上对自动分类的研究主要着重于对互联网资源的自动分类和标引。

每个数字信息对象是否被分配了唯一的识别符或名称?数字对象由两部分构成:内容和元数据。一个数字对象可含有多种类型的内容,如文字、图像、音频。复杂的数据类型可由简单类型构造。元数据包括句柄、权限、访问方法、数字签名、交易日志等。数字化信息对象只有被赋予唯一的标识符或名称,才可能被识别和检索。名称和标识符是数字图书馆的基础建构块。名称用于标识数字对象,注册数字对象中的知识产权、记录所有权的变化,在引用、检索和对象链接中不可缺少。名称必须是唯一的,必须长期保持。数字对象的唯一标识符——句柄(计算机系统可以通过句柄来唯一地指示文件,句柄就是一种指针)是用于标识数字对象的唯一的一个字串,独立于其存储位置且长期有效。例如,William Y. Arms于1997年2月在Corporation for National Research Initiatives(CNRI)出版的D-lib Magazine上发表的著名的An Architecture for Information in Digital Libraries一文的句柄可表示为:hdl:cnri.dlib/february97-arms。

3) 知识组织体系的建立

从网络信息资源组织的角度看, 现有各种知识组织体系分为三个层次, 它们在结构、复杂性、功能等方面存在明显的差异。第一个层次是词汇表, 强调概念的定义, 一般不涉及复杂语义关系和分类结构, 如规范档 (Authority Files)、术语表和字典等; 第二个层次是分类体系, 强调概念间的层次、聚类 and 类别体系, 如主题词表和分类表; 第三个层次是关系列表, 强调对概念之间关系的表现, 如叙词表 (Thesauri)、语义网络 (Semantic Networks) 和本体 (Ontology)。

数字图书馆是否提供这三个层次的知识组织体系? 是否利用知识地图 (Knowledge Map) 等手段实现这三个层次的可视化知识组织?

4) 海量信息的集成

是否能够较好地实现海量信息的存储? 如何保存和管理海量数据是数字图书馆系统设计的核心任务之一。海量信息的存储需要大容量和高密度的电子存储设备, 包括大容量、高密度的硬盘和光盘及光盘塔和光盘库, 对那些馆藏达几百万册, 以至上千万册的大型图书馆来说, 对存储设备的容量要求更高。

海量信息的集成系统主要由信息服务器、信息检索部件、查询处理单元、数据仓库、知识库、集成平台与工具、异构信息源 7 部分构成。好的数字图书馆系统应能将各种数据库进行整合和提供应用并具备文件管理、文件编辑、版权管理、内容管理、多媒体管理、信息输出和发布等功能。

5) 海量数据的搜索

信息组织的目的是为了提供检索, 海量数据的搜索效率 (包括多语种搜索、图像搜索、语音搜索、智能搜索) 与速度是衡量数字图书馆信息组织效果的重要指标。

能否开发出一个比较好的检索工具来实现对海量信息仓储的检索? 一个比较好的检索工具能够使得提供给用户的信息恰恰是用户最需要的 (不需要的全部剔除), 要达到这样的要求, 就要采用异构信息的专用搜索引擎组, 而目前的一些搜索引擎还达不到这样的要求。

能否支持大量并发用户的同时访问?

是否采用了可视化的信息资源描述与组织方法, 使得用户的检索操作可以实现可视化? 现有的搜索引擎的信息检索过程是“黑箱”的, 用户无法控制, 也无法实现对检索结果总体情况的了解。数字图书馆必须考虑用户的特点与信息行为, 可视化信息检索是数字图书馆信息检索的发展方向之一。

是否提供智能化的信息访问? 数字图书馆信息组织的基本目标是创造一个良好的信息环境, 提供对分布式存储的信息的知识化组织、智能化访问和服务。智能化访问是指对信息的访问不仅仅简单地对原始数据的查找 (如当前 Web 搜索引擎的关键词查找), 而是根据用户的信息需求进行知识查找和内容提取。

6) 信息的呈现

数字图书馆的信息是否利用信息可视化技术、虚拟现实等技术, 通过图像、图形、语音等直觉化手段表现出来? (直觉化的表现手段有助于研究人员从中汲取许多新的思想, 也有利于计算机从这些数据中得到更重要的知识。)

对用户检索出的信息是否以图、文、声方式进行综合显现? 一个可视化的环境可以为用户展示更丰富、更直观的信息。

对系统环境、解读软件系统的依赖程度如何? 硬件、软件的更新换代是否会影响数字信息的长期读取? 现有的多媒体图形的表现是否可以进一步细化?

对多媒体信息能否实现分层传输? 在分层传输中, 如果用户提交了一个多媒体信息对象的请求, 并能一次找到很多照片或图像, 系统将可以找到的照片分成若干层, 将最粗的那层

传给用户，用户确认选择后，再逐渐细化，而当用户认为这张照片并非所要的内容时，可随时结束，再换另一张。能否克服多语言问题（包括机器翻译问题、多语言浏览器问题）？能否仔细区分生成的原始数字对象、存储在仓储中的数字对象和提交给用户的数字对象？数字对象使用时和存储时的形态可能完全不同。图像可以以一组小波存储在图书馆中，利用时再根据要求用小波生成图像。例如，可将音乐作品的曲谱直接传给用户，也可在仓储中用合成器演奏后将音频信号传给用户。

7) 信息安全

信息数字化后，给用户利用提供了极大的方便，但同时也使非法盗取、复制、修改他人作品变得轻而易举；硬件与操作系统的更新给数字信息的长期存取带来了困难；数据的意外丢失、计算机病毒和“黑客”也时刻危及着数字图书馆的信息安全。因此，信息的安全保护问题是数字图书馆信息组织过程中要考虑的另一个重要问题。

数字图书馆的信息安全是否有保障？

是否采用各种技术来保护数字图书馆信息安全？例如，可以利用数字水印技术（在图像中明显地或隐藏地标记数字知识产权信息）控制网上数字信息的完整性，预防黑客对数字信息的涂改，而利用“电子声音技术”实施电子追踪，版权人可以通过它来搜索作品被利用的信息，控制作品的访问利用。

仓储能否保管好所持有的信息？数字对象包含有价值的知识产权，因此，仓储中的数字对象含有是否允许其在特定的经济和社会框架中被操作的信息（保存在元数据中）。仓储必须妥善管理，提供参考引用、安全控制等措施，以确保对数字对象的操作是合法的。用户通过仓储访问协议与仓储交互，以屏蔽仓储的内部结构和数字对象的存储形式。协议中的命令包括访问数字对象及其元数据、添加和删除、传递请求等。

8) 用户界面

用户界面是数字图书馆的重要组成部分，是系统展现在用户面前的窗口。数字图书馆的信息要被广大用户利用，信息组织中的接口设计就要充分利用图形、语音等手段，让用户使用时得心应手，能够根据自己的要求进行个性化的定制（Personalized Customization），并具有人性化、智能化的特性。

是否设计了一个用户满意、易用、高效的界面？

是否提供了用户自我服务系统和请求帮助系统？这是数字化图书馆的重要组成部分，前者能在客户端上显示读者指南，能自动指引读者使用数字化图书馆。目前大多数电子信息中心均有自我服务系统。请求帮助系统能在用户不中断检索的情况下，一步一步地帮助用户解决问题；系统专家还能监控这些活动，知道信息专家解决问题的情况。数字化图书馆应有各种信息专家，随时接受读者的联机访问并提供咨询。

3. “美国记忆”（American Memory）数字图书馆项目评析

“美国记忆”（American Memory）即美国国会图书馆正式启动的美国国家数字图书馆项目（National Digital Library Program, NDLP），是一个大规模的数字图书馆项目，将美国国会图书馆的历史馆藏电子化，并通过 Internet 提供访问。该项目旨在让所有的学校、图书馆、家庭同那些公共阅览室的长期读者一样，能够在所在地便捷地接触到这些对他们来说崭新而重要的资料，并按个人要求理解、重新整理和使用这些资料。

由于该数字图书馆项目建设的内容是国会图书馆的有关美国历史的馆藏，因此它采集的信息组织对象在内容上相对单一。迄今，“美国记忆”储存了超过 900 万条有关美国历史和文化信息的数字化资源记录，它们大部分来源于国会图书馆丰富的馆藏，包括文本、手稿、照片、地图、乐谱、音频及视频文件等形式，并以图、文、声的方式进行综合显现。同时，

国会图书馆还充分利用现有的数字化资源,它在1996—1999年期间联合美国科技公司开展“全国数字图书馆竞赛”,吸引其他图书馆、科研机构、博物馆等单位提供它们自身的数字化资源,进一步丰富了“美国记忆”的资源总量。这些机构提供的信息资源不仅翔实、权威,同时也是比较固定的信息源。“美国记忆”信息采集的范围在形式上包括网上资源的下载和对馆藏文献的数字化。前者通过资源下载,组织和整合后提供给用户使用。后者使用扫描仪、数码相机等将视听资料数字化或人工输入,将大量文献资料、珍贵手稿、档案、实物转换成数字化信息。

“美国记忆”的信息加工对象是不同媒体、不同格式的信息资源,除文字信息外还包括声音、图像、视频等类型的信息。因此在信息加工中该项目采用国家标准或成熟的行业标准,如图像先保存为未压缩的.tiff文件,使用JPEG算法压缩,再转换成适于网络使用的.gif格式;如果没有成熟的标准,则利用当时刚出现的格式,如地图用MrSid格式,活动图像采用QuickTime格式,视频采用Real Audio格式。

“美国记忆”在建设初期并没有统一的数字化资源描述标准,且多种资源描述格式的转换存在一定障碍,限制了资源共享。鉴于此,1996年国会图书馆要求参与“全国数字图书馆竞赛”的机构在数字化资源时统一采用由它规定的标准。2001年年初,国会图书馆开始对“美国记忆”资源格式进行全面转换,以实现两种主要的资源描述格式:XML schema中的MARC格式和都柏林元数据核心元素集,从而开始利用元数据的多个元素对信息资源进行多维揭示。随着开放存取运动的兴起,它采用OAI-PMH协议实现多种资源的整合。

“美国记忆”在一定程度上对其信息资源做了非线性组织。但相关联的信息点只局限在主题词和作者名称上,用户通过点击一条记录的主题词和作者名称链接到与之相关的记录上。它采用句柄来为数字对象提供一个唯一的、永久的、独立于位置的名字供用户识别。

它还注意将人工与自动处理相结合。例如,一幅扫描图片的保存入库过程如下:选择待转换为数字对象的资料(人工);在相应的字段中给定元数据值(人工);建立元对象及其指向其他数字对象的链接(自动);将数字对象置入仓储中(自动);在句柄系统中注册句柄(自动)。

“美国记忆”采用主题词表和分类表的形式建立其知识组织体系。用户可据此进行主题浏览(Browse by Topic),它的一级类目(按类目英文名称字顺排列)分为广告、非洲裔美国人史、建筑与风景、城市和城镇、民间文化、环境与环境保护、政府和法律、移民和美国扩张、文学、地图、本土美国人史、表演艺术与音乐、总统、宗教、运动休闲、工业技术、军事战争、妇女史18大类,每个一级类目下再根据该类目下的信息资源特点按照不同划分标准分成各二级类目。

为了方便用户从不同角度浏览资源,“美国记忆”还提供了其他浏览途径,如按时间浏览(Browse Collections by Time Period)、按文献类型浏览(Browse Collections Containing)和按地点浏览(Browse Collections by Place)。

“美国记忆”采用分类索引和搜索引擎两种检索方式。和国会图书馆的其他一些项目一样,“美国记忆”使用了智能信息检索中心(Center for Intelligent Information Retrieval, IIRC)研制的一种索引和检索引擎InQuery来实现这两种功能。在分类索引方面,“美国记忆”提供主题、时间、地点及格式等多个检索入口,方便用户从不同角度查找信息。其分类目录较为详尽,在首页各大类下提供二级甚至三级类目,方便用户迅速了解资源库的范围和内容。它还对每一个记录进行简要描述,使用户在浏览之前对其能有更全面的认识。

InQuery的检索功能非常强大,提供了15种以上的查询操作,其中既包括严格的逻辑操作、邻近操作,还提供了基于概率模型的一些操作。其检索的灵活性也较强,允许用户选择

不同的检索途径，既可在全部资源或某一特定主题数据库中进行搜索，也可通过主题和题名分类查找，同时，还可以在搜索结果中进行二次检索，并且还允许用户给查询表达式中的检索词赋予不同的权重，进一步确保了检准率。

InQuery 还提供了三种检索策略供用户选择，即检索结果可包含检索式中任何一个字符，或包含全部字符，或包含完整字句。在后台，InQuery 会分别检索书目记录索引和全文索引，如果查询表达式中有多个检索词，这种搜索动作会进行很多次，并将几次搜索的结果合并到一起。最后返回的命中文献有以下几种情况：当查询表达式中检索词构成的短语与文献中的一个短语完全一致时，该文献将排在命中文献列表的前面；当查询表达式中所有检索词均在文献中出现，此时这几个检索词在文献中的距离将决定该文献在命中列表中的位置，距离较近的排在前面；如果上面两种情况都不存在，那些仅包含查询表达式中部分检索词的文献也将作为命中文献返回。此外，其检索结果的显示也允许用户按照相关度来排列。

在知识产权与隐私权的保护上，“美国记忆”设立了专门的“法律声明”网页，就网站安全、知识产权、隐私权等内容加以详细说明。在网站安全方面，国会图书馆采用 Cookie 追踪软件来监视“美国记忆”网站的访问情况，以识别那些未经授权就上载或修改网站资源信息、破坏网站服务的行为。在知识产权保护方面，国会图书馆通过各种途径与著作权人联系，争取到著作权人或与著作权人有关的权利人的授权，并且提请尚未联系上的著作权人主动与其联系。在每一个主题数据库的页面上，都特别注明该资源是否有知识产权方面的限制及限制程度。在隐私权保护方面，国会图书馆明确说明了用户在访问“美国记忆”网站过程中，图书馆将收集存储的信息，包括用户访问互联网的域名、访问时间、访问过的网页和文档等。但图书馆将尊重用户的隐私，如果用户将其个人信息提供给国会图书馆，在得到用户允许以前，该馆不会将这些信息透露给其他组织或个人，并向用户说明该网站使用了 Cookie 追踪软件跟踪分析用户的网上行为。此外，“美国记忆”在其每一个网页上都提供了“法律声明”的链接，随时提醒用户注意相关的法律条文。

“美国记忆”的界面简洁易用，考虑到核心用户是中小学教师和学生，“美国记忆”的搜索界面上只有一个输入窗口，这样，用户无须知道系统的这些细节，无须培训就能使用。主页上列出了一级大类，便于用户浏览检索。

它的帮助系统也较为完善，包括如何浏览音频/视频文件、检索帮助、FAQ 和在线帮助。用户可通过“Ask a Librarian”提交表单或通过“Chat with a Librarian”实时在线参考咨询的形式向国会图书馆的图书馆员和“美国记忆”信息专家提问以寻求帮助。

从上述的分析中不难看出，“美国记忆”在信息采集、组织加工、检索功能设置、安全保护及帮助系统方面都做出了有益的探索。但是，在交互性、智能性、信息推送和个性化服务定制等方面显得明显不足。

总之，虽然数字图书馆尚处于探索阶段，但在现有的各种网络信息组织模式中，它最有发展前景，将成为下一代互联网网上信息资源的组织与管理模式。



本章小结

本章主要论述了非网络环境下和网络环境下检索工具或系统信息组织的实例与评价。列举了印刷型文献、多媒体信息、网站、搜索引擎、学科信息门户、Web2.0 网络应用以及数字图书馆信息组织的成果实例，提出了各种信息组织方式的评价标准，并对其价值、优劣等进行评价。



问题讨论

1. 印刷型文献信息组织、光盘数据库信息组织和网络信息资源组织各有什么特点?
2. 印刷型文献信息组织、光盘数据库信息组织和网络信息资源组织的评价标准有何异同?
3. 光盘数据库与网络版数据库信息组织有何异同?
4. 搜索引擎信息资源组织的评价指标有哪些?
5. 学科信息门户信息资源组织的评价指标有哪些?
6. Web2.0 环境下信息资源组织的主要方式有哪些,各自的评价指标是什么?
7. 数字图书馆信息组织的评价标准有哪些?



内 容 索 引

说明: ① 本索引以书中内容关键词为标目, 以标目词被重点论述部分的页码为出处; ② 本索引共分三部分, 中、外文标目分别编排, 阿拉伯数字开头的索引排在最后; ③ 中文部分按标目汉语拼音音序排列, 外文部分按标目逐词排列, 在各字头前加首字母以增加助记和助检功能; ④ 人名、书名索引标引与其他关键词索引混排, 书名索引标目加书名号, 书名号不排序。

一、中文索引

A

艾奇逊 25
爱因斯坦 53
安全技术 390
奥斯汀 78
奥特勒 19, 232

B

八分法 224
百度 27, 383
百分法 224
百科全书 256, 412, 416, 418
百科全书派 60
柏拉图 50
版本 357
版本目录 354
版本信息 115
半自动标引 307
包 103
包含关系 57
保存 170
保存型元数据 170
保留上下文索引 78
保留上下文索引系统 24
报刊全文数据库 385
报纸目录 355
报纸索引 363

贝塔朗菲 40
被控制词 89
被引率 435
被引用次数 366
被引著者 365
本体 29, 90, 91, 92, 93, 207, 238
本体表示 95
本体表示语言 94
本体分析 95
本体工程 94, 95
本体构建 94
本体互换语言 110
本体推理层 110
本体论 50, 90
本体评价 95
比较功能 80
比较关系 289
笔画笔形法 329
编辑注释 119
编码偏好程度最小 94
《编目服务通报》 267
编目索引 363
编年排检法 341
编制说明 214
变更注释 119
标记符号 205, 214, 218

标记制度 221
标目 134
标目法 140
标谱 412
标签 101
标签关系 121
标签集 101
标签相关 121
标识 288, 291
标识符 136, 237
标识符号 140
标题 291
《标题表》 21
标题词 81, 288
标题词语言 54
标题法 54, 81, 206
标题款目 268
标题语言 288
标引 246
标引不控制 88
标引词 292
标引词加权方法 308
标引法 246
标引工具 288
标引功能 79
《标引和分类学语言——结构和功能的语言学研究》 55

标引控制 88
标引语言 288
标志信息 199
标准复分表 213
标准化 23, 33
《标准机器能读目录款项的建议》 141
标准通用标记语言 98
表达性 219
表示本体 51
表述方式 138
别裁 17
《别录》 354
并列关系 57
波普尔 48
玻耳兹曼 41
玻耳兹曼方程 42
博科 26
不标引 85, 87
《不列颠百科全书》 421
不相交类 112
不相容关系 57
布尔组合 114
《布立斯书目分类法》 209
布利斯 24, 49
布鲁克斯 49
部件词典法 313
部首查字法 326
部首法 327
部首排检法 326
部首相关 112

C

材料复分表 213
采集器 382
参见 261
参见自 261
参考文献 366
参考信息源 126
参考元数据 170
参照 134

参照项 268, 282
层 239
层累标记制 221
查修 21
查字法 322
《常规武器分面叙词表》 25
《常规武器专业主题词表》 25
超级组配 294
超文本 350
超文本标记语言 98
超文本方法 10
超文本链接 350
超星数字图书馆 395
陈述 188
陈思 17
程长源 24
《出版机构与国家书目服务连接》 34
储存与压缩技术 389
触摸屏输入 32
传统著录法 124
垂直搜索引擎 380
词表标记语言 100
词典 256, 412
词典切分标引法 313, 316
词典实现模式 90
词汇标签 118
词汇控制 75
词汇选择 75
词间关系的控制 76
词频统计标引法 308
词频统计模式 90
词区分值加权标引法 309
词相关性加权标引法 310
词形的控制 76
词形实现模式 90
词义的控制 76
词族索引 273
《辞海》 50
次要主题 289
从页 436

丛书 255, 357
丛书目录 355

D

《大英百科全书》 421
代 265
代号 218
代码排序法 349
代码语言 54, 82
戴维民 55
戴志骞 21
单纯号码 219
单词语法 80
单个主题词 283
单汉字标引法 314, 316
单语种搜索引擎 380
单元词法 54, 81
单元词卡系统 78
单元词语言 54
单元主题 289
单主题 253, 288
《档案著录规则》 136
导航 439
倒置标题 260
倒转培根分类法 61
登记列举式分类法 206
等级标记制 221
等级关系 261
等级聚类法 301
等级-组式分类语言 54
等同关键 89
等同类 112
狄德罗 19, 40, 59
迪昂 47
迪昂-纽拉特-蒯因论题 47
地点和频率法 340
地理复分 269
地区复分表 213, 262
地毯曲线 52
地图检索 423
地图索引 423

《地图资料著录规则》 136

地序法 348

地序排检法 348

地序组织法 12

地支纪时法 346

地址目次区 146

地址目次区 151

帝王年号纪年法 343

电子杜威 228

《电子技术汉语主题词表》 24

调用等级聚类法 302

迭代生成原则 52

顶层元素 181

顶级概念 117

定位 170

定义 119

定义层 110

定义域 107, 108

定组式标题 262

动力 207

动态聚类法 302

动因 207

都柏林核心倡议 171

都柏林核心集 100

都柏林核心元数据 26

都柏林核心元素集 171

独立搜索引擎 379

杜邦公司 22, 78

杜定友 21

杜威 19, 61, 228

《杜威法》 216

《杜威十进分类法》 19, 64,

209, 228

对称属性 114

对口标引 287

对象层 110

对象属性 113

对应编号法 227

多卷书 255

多语种搜索引擎 380

多主题 254, 289

E

恩格斯 19, 60, 61

二次信息 9

二分法 58

《二十五史人名索引》 21

F

发行目录 355

发行信息 199

《法国百科全书》 19

反对关系 57

反函数属性 114

泛指词 261

范畴职能 77

范畴职能引用次序 290

范例 119

范围注释 119

方志目录 355

仿分 213, 251

非标题款目 269

非关键词表 299

非记录型信息 10

非受控主题法 267

非书资料 258

非书资料目录 355

《非书资料著录规则》 136

非数字信息 9

非印刷型资料 258

非用词后缀表法 312

非正式标题 269

非正式主题词 272

非叙词 273

非纸质资料 258

非专业主题 289

《吠陀经》 17

分布式共享 51

分解转换 290

分类 204, 301

分类标引 246, 252, 297

分类标引规则 252

分类表 276

分类表索引 216

分类法 204, 297

分类号 218, 268

《分类号-主题词对应表》 280

分类检索工具 205

分类索引 366

分类统计 205

分类语言 54, 80

分类语言 80

分类主题词 266

分类主题一体化语言 266

分类自动标引 86

分类组织法 10

分面标记制 221

分面公式 239

《分面叙词表》 25, 266

分面组配式分类法 206

分面组配式分类语言 54

分散标引 256

分维 52

分析标引 287

分析器 382

分形维 52

分型理论 52

服务点播 29

辅助表 212

辅助符号 137

辅助索引 273

负熵原理 48

负载元数据 170

附录 130, 134, 212, 273

复分 215, 251, 269

复分标题 260, 268

复分表 212

复分顺序 270

复合主题 289

复杂性 110

副表 212

赋词标引 307

覆盖范围 169

G

概率熵 48
概率组织法 13
概念 92, 110, 117
概念地图 29
概念分解 291
概念化 91
概念体系 118
概念映射关系 121
概念组配 294
概述 130
干支纪年法 344
高等教育文献保障系统 27
《高等协同学》 43
戈尔曼 132
格鲁伯 91
格斯纳 17
《各科著名学者及其著作目录》
17
根元素 181
更新频率与滞后周期 387
公理 92
功能性 110
共享 91
共性区分表 212
古代纪日法 345
古代纪时法 346
古代纪月法 345
古代信息组织 15
古典书 354
古籍 356
《古籍著录规则》 136
《古越藏书楼目录》 20
谷歌 27, 383
骨架法 95
关键词 82, 297, 300
关键词标引 297
关键词标引法 246
关键词法 54, 81, 297
关键词语言 54
关联 101

关联符号 263
关系符号 78
《关于信息构建的 10 个问题》
45
观察陈述 47
官名索引 349
官修书目 354
管理 170
管理型元数据 170
光盘数据库 430
广义词 261
归类 246, 301
归约 110
规定复分表 283
规定信息源 127
规范格式 389
规范语言 74, 83
郭兰克钮 50
《国防科学技术叙词表》 96
《国防科学技术主题词典》 25
《国会标题表》 231
《国会法》 219
《国会图书馆分类法索引》 236
《国会图书馆分类法索引汇编》
236
《国际编目与书目控制》 34
国际标准化组织 33, 134
国际标准刊号 134
《国际标准书目著录》 128, 134
《国际分类》 49
国际教育标准分类 67
《国际十进分类法》 19, 232
国际图联 33, 142
国际性原则 174
国际知识组织协会 33
国家书目 354

H

哈肯 43
哈利 61
哈钦斯 55

函数 92
函数属性 114
《汉表》 271, 280, 285
《汉语拼音方案》 323
汉语拼音检字法 323
汉语信息自动标引 311
《汉语主题词表》 25, 271, 284,
汉字拼音排检法 323
汉字信息处理工程 271
《汉字信息处理系统工程》 25
汉字形序排检法 326
汉字音序排检法 322
《航空科技资料主题表》 24
豪斯道夫 53
号码法 330
号码配置 250
号码索引 367
耗散结构理论 42
合适取值原则 174
何多源 21
核心本体 51
核心出版物 387
核心类 107
核心属性 108
赫尔姆 210
黑格尔 50, 60
洪业 21, 331
后控制 88
后控制词表 88
后组式 82
后组式主题法 267
互著 17
《化学文摘》 2, 81, 298
环境复分表 213
《皇览》 17
黄邦先 17
黄纯元 49
会议录索引 364
混合号码 219

J

机读目录 141
机读目录著录法 124
机检词 292
机器标记语言 97
机器可读目录 141
《机械工程主题词表》 25
积累提问式模式 90
基本部类 209
基本大类 209
基本分面 276
基本科学指标 432
《基础叙词表》 25
基数约束 113
基于 XML 的本体交换语言
100
疾病标题复分表 262
集成词表 266
集成搜索引擎 380
集合 103
集合成员 121
集合成员列表 121
集中标引 256
计算机辅助标引 306
计算机检索技术 390
计算完备性 112
计算语言学 55
记录 125
记录次序 138
记录头标区 150
纪年排检法 341
技术标准 258
技术元数据 170
加菲尔德 84
加权 78
加权标引 317
检索 170, 439
检索不控制 88
检索点目标 125
检索结果分组 394
检索控制 88

检索用词表 88
检索语言 53
简表 211
简单易用性原则 173
简单知识组织系统 116
简明性 220
建模元语 100
建模元语 109
《建设工业叙词表》 25
交叉关系 57
交叉组配 294
交替类目 215
角色 110
《教育分面叙词表》 25
《教育杂志》 20
《教育资源叙词表》 82
揭示功能 79
节点 350
节目 354
结构型元数据 170
借号法 225
近代信息组织 18
禁用词 82
禁用词表 299
句法独立原则 173
句法控制 77
句式变换 79
聚类 300
聚类控制实现模式 90
绝对词频加权法 308

K

卡利马科斯 17
卡片式目录 356
开发性知识库的链接协议 100
开放的统一资源定位器 400
开明书店 21
凯赛 77
康德 40
抗体系统理论 40
靠词标引 294

科茨 77
科技报告 258
科契 52
《科图法》 209, 222
《科学》 52
《科学分类之历史》 21
《科学引文索引》 82, 370
可变长字段 147
可复性 173
可扩长性 94
可扩展标记语言 99
可扩展性原则 173
可视化 32
可视化信息检索 318
可修改性 173
可选标签 118
可选择性 173
克劳修斯 41
克里斯塔多罗 20
克特 19, 77
克特号书号 236
克特著者号 236
空间 207, 239
空间参考信息 199
空间数据组织信息 199
控制词 89
控制符号 78
控制字段 147
跨媒体搜索 318
蒯因 47
款目 125
框架模式 100
奎因 47
扩充性 219
扩九法 224

L

拉芳丹 232
拉封丹 19
来源著者 365
乐谱目录 355

类 92, 107, 112

类间组配 251

类目参照 215

类目分析 217

类目索引 216

类目体系 217

类目注释 249

类书 412

类缘关系 265

历法常识 341

历史注释 119

《连续出版物著录规则》 136

连续划分 58

联邦地理数据委员会 198

联号 292

《联合国教科文组织分类法》
62

联合国教科文组织 142

联机编目 23

联机公共检索目录 356

联机公共目录查询系统 26

联机计算机图书馆中心 23

联结主题 289

联系符号 78

联系信息 200

链 350

链接 112, 438

梁启超 20

列举 115

列举式书目 354

列举-组配式分类法 208

列序法 334

林保 17

林耐学派 60

领域本体 51

刘国钧 20

刘湘生 78

刘向 16, 354

刘歆 16

卢恩 22, 25

吕绍虞 21

《伦敦教育分类法 (第二版)》
25

轮排索引 273

论题复分 269

逻辑形式 109

逻辑学 55

M

马丁·凯 55

马张华 59

麦尔威·杜威 64

曼德布罗特 52

毛泽东 60

矛盾关系 57

《冒号分类法》 21, 207, 237

《美国出版商目录年报》 359

美国国防部文献保障中心 26

《美国国会图书馆标题表》 81,
267

《美国国会图书馆分类法》 19

美国国会图书馆联机目录 362

《美国国会图书馆图书分类法》
210, 234

《美国国会图书馆字典目录用
标题表》 267

美国海军兵器中心 21

美国记忆 458

美国剑桥科学文献社 375

美国全国联合目录 361

美国生物医学本体中心 100

《美国图书馆协会标题表》 20

《美国在版书目: 著者、书目》
359

《美国在版书目主题指南》 359

美国专利局 21

美加连续出版物联合目录 362

米尔斯 24

宓浩 49

密度法 302

描述 170

描述逻辑 110

描述数据 169

描述项目 125

描述型元数据 170

描述性书目 354

描述著录 134

名录 412

明兰茂著 325

命名空间 100

墨子 17

目录 354, 412

目录式搜索引擎 377, 381

目录之目录 12

穆尔斯 24

N

内容标识 6

内容索引 363

内容特征 204

内容元数据 170

内在性原则 173

奈斯比特 41

能量 239

逆文献频率加权标引法 309

年代复分 269

年代排检法 341

年鉴 256, 412

牛顿 60

纽拉特 47

农业本体 96

农业叙词表 96

P

排列次序 209

判定性 112

培根 59

配号方法 215

皮高品 20

《皮纳克斯》 17

拼音法 323

平均信息量 42

评估 170

评论性书目 354

普尔 20
《普尔期刊文摘索引》 20
普利高津 42
《普通高等学校本科专业目录》
63
《普通图书著录规则》 136
谱序法 349

Q

七步法 95
《七录》 16
《七略》 16
《七志》 16
期刊目录 355
期刊索引 363
期刊文献读者指南 373
期刊引证报告 432
齐夫 306
齐夫定律 306
前控制词表 88
钱学森 61
浅标引 286
切分标引法 311, 316
倾向关系 289
《清代文集编目索引》 21
《情报检索语言语法体系初探》
55
庆应义塾大学 26
圈码 283
全标引 85
《全国报刊索引》 367
全国信息技术标准委员会 34
全面标引 286
全球书目控制和国际机读目录
核心活动 34
全球万维网联盟 99
全文检索 84
全文链接 434
全文收录覆盖时间 387
全文数据库 385
全文搜索引擎 378

全文索引 171
全信息 47
全自动标引 307
权值排序法 349
权重值信息组织法 13
诠释数据 169
《群史姓纂歆谱》 17

R

《人大法》 226
人工神经网络 28
人工语言 74
人工智能 306
人工智能模式 90
人工智能研究中心 100
《人民大学图书馆图书分类法》
226
人气质量定律 341
人文科学索引 373
人文社会科学 68
人物标题复分表 262
人物表 213
人物传记索引 374
任务本体 92
容纳性 219
容器 103
容器层 110
容器元素 192
如果必要 135
如果合适 135
阮冈纳赞 21, 78, 207, 237
阮孝绪 16

S

赛伯空间 49
赛格尔 55
三次信息 9
商业信息、统计报告全文数据库 386
熵 41
上位词标引 293
上位类 247

上位类传递关系 119
上位匹配 121
上下文关键词索引 298
社会科学 68
《社会科学检索词表》 276
社会科学文献索引 373
社会信息 3
深标引 286
神经网络分词法 315
审核 288
《生命是什么?》 42
生物本体核心组织 100
生肖纪年法 345
声部排检法 326
声明 107
声音识别 32
声韵法 325
《圣经》 17
圣西门 60
《十三经索引》 21
时代复分表 213, 262
时间 207, 239
时间段信息 200
时间因素 290
时序法 341, 347
时序排检法 341
时序组织法 12
实体和属性信息 199
实用分类系统 90
史志目录 354
使用范围 140
使用型元数据 170
世界 348
《世界书目》 17
《世界图书总目手册》 232
世界种族与民族复分表 213
事物性质分类排检法 336
视窗杜威 228
视频信息 8
视频信息标引 319
收藏目录 355

首选标签 118
受控词表 116
受控语言 74
受控主题法 267
《授予博士、硕士学位和培养研究生的学科、专业目录》 65
书本式目录 356
书目 355, 412
《书目分类法》 24
书目著录 124
书评索引 364
书评摘要 374
术语 100
数据 2
数据标引 389
数据加载 389
数据检索 389
数据库导航 402
数据库技术 389
数据库检索系统 387
数据类型 115
数据维护 389
数据系统 2
数据质量信息 199
数据属性 113
数据准备 388
数据字段 147
数序纪日法 346
数字化地理元数据的内容标准 198
数字图书馆 453
数字信息 10
数字信息组织 10
数字园区 44
双表列举 242
双位加点法 226
双位制 224
双重关键词索引 299
顺排检法 322
顺序标记制 221
顺序-层累标记制 221

说明 214
《说文解字》 327
私撰书目 354
斯巴克·琼斯 55
《四角号码笔形代码表》 330
四角号码法 330
《四库全书总目》 16
搜索引擎 27, 340, 377, 445
《隋书·经籍志》 16
岁星纪年法 344
孙毓修 20
缩微目录 356
索引 257, 363
索引学会 20
《索引与索引法》 21

T

《台湾人文学学科引文索引》 366
《台湾社会科学引文索引》 366
太岁纪年法 344
太阳历 342
太阴历 342
陶伯 21, 24
陶布 54
特性检索 10
特种文献 258
题目索引 366
题内关键词索引 298
题外关键词索引 298
体系分类排检法 335
体系式分类语言 54
体系限定词 175
替代机制 6
替代记录 5
替换 103
天色纪时法 346
通用本体 51, 92
通用复分表 262
通用因素 290
《通志·文艺略》 16

《通志·校雠略》 17
同被引 366
同为类 248
同一关系 56
同义词 89
统计标引法 306
统一导航 394
统一题名 134
头标区 144
图录 412
《图书馆》 20
《图书馆编目技术》 20
《图书馆图书、小册子排架和编目用分类法及主题索引》 228
图书目录 355
图书情报知识论 48
图像 8
图像信息 8

W

外文目录举要 359
外文字顺法 332
万方数据库资源系统 395
万维网 91
万维网杜威分类法 27
万维网联盟 34, 98
万维网虚拟图书馆 27
王公在位纪年法 343
王俭 16
王永成 22
王云五 330
王重民 21
王子舟 49
网格计算 36
网络版数据库 430
网络本体语言 111
网络二次信息 11
网络机器人 (robot) 382
网络技术 29
网络内容选择

- 网络爬虫 (crawler) 382
网络三次信息 11
网络数据库 432
网络信息分类排检法 337
网络信息关键词排检法 339
网络信息组织 10
网络一次信息 10
网络知识组织 31
网络知识组织系统 97
网络蜘蛛 (spider) 382
网络资源 258
网络资源指南 444
网站 112, 436
网站述评 435
网址检索 181
威尔逊精选全文数据库 393
威妥玛 326
威妥玛式拼音排检法 326
维克利 78
位置因素 290
谓词 102
魏征 16
温克勒 132
文本分类 303
文本预处理 389
文档类型定义 99
文科 68
文献 4
文献保证 222
文献交流论 49
文献类型复分表 262
文献类型因素 290
《文献目录信息交换用磁带格式》 143
《文献目录信息交换用磁带记录格式》 141
文献著录标准 23
《文献著录总则》 124, 136
文摘 257, 363
《文摘编写规则》 376
文字 8
文字识别 32
沃尔夫 50
沃尔曼 15, 43
五笔字型法 331
物质 207, 239
- X**
- 西班牙 34
西文全文数据库举要 390
《西学书目表》 20
系统论 40
狭义词 261
下级标题词 261
下位类 247
下位类关系 119
下位匹配 121
先控制 88
先组定组式主题法 266
先组散组式主题法 266
先组式 81
显性主题 289
显著特征 204
显著性 77
显著性引用次序 290
《现代汉语词典》 415
现代书目 354
相关关系 100, 119, 261, 265
相关记录 434
相关匹配 121
相关索引 216
相关文献 366
相关主题 289
相容 110
香农 41, 48
详表 211
详简级次 127, 140
向量空间模型
向上兼容性原则 173
《小字录》 17
《校讎通义》 17
协同学 43
《协同学导论》 43
谢尔·凯门 45
谢尔宾斯基 52
《新版中国机读目录格式使用手册》 143
新浪 440
《歆海镜源》 16
信息 4
信息安全 458
信息爆炸 45
信息标引 5
信息采集 382
信息储存 7
信息分析 7
信息符号 15
信息构成 169
信息构建 29, 43
信息构建师 45
信息构建研究会 44
信息管理 2
信息饥渴 45
信息集合 44
信息检索 93
信息交换论坛 198
信息焦虑 45
信息结构 44
信息可视化 30
信息空间 45
信息描述 7, 74, 93
信息片段 44
信息熵 42
信息社会标准化系统 34
信息特征 74
信息体 2
信息挖掘 33
信息网关 406
信息系统 2
信息序化 5
信息选择 7
信息有序化 6
信息语言 55

信息源 7, 126, 140

信息载体 4

信息增值 2

信息整理 4

信息整序 3

信息著录 5

信息资源 2

信息资源加工技术 389

信息资源组织 3

信息组织 2, 6, 14, 55, 382

信息组织空间 3

信息组织语义网格化 36

信息组织智能化 36

形式复分 269

形式化 91

形序法 322

虚拟图书馆 27

虚拟现实 49

许慎 327

序 4, 103

序论 214

叙词 81, 264, 288

叙词标引 293

叙词表 82, 288

叙词法 54, 81, 96, 264

叙词化 278

叙词语言 54, 264, 288

叙词组配 294

选词标引 293

选择 170

选择性规则

薛定谔 42

学科导航 27

《学科分类与代码》 62, 72

学科体系分类排检法 334

学科信息门户 406

学位论文 258

雪花曲线 52

巡 239

Y

雅虎 32

亚里士多德 17, 59

亚历山大图书馆 17

亚述巴尼拔 17

研究活动 435

颜真卿 16

衍生复分 269

阳历 342

杨志远 382

姚明达 16

《药学总览》 20

业务标注 125

叶圣陶 21

一般分类规则 253

一步到位的服务 29

一次划分 58

一次信息 8

一体化词表 266

艺术索引 374

《艺术与建筑叙词表》 96

艺术作品描述目录 194

异构数据库整合 395, 401

逸馥 62

因果关系 289

阴历 342

阴阳历 342

音频信息 8

音序法 322

《引得说》 21

引文被引文献 365

引文链接 434

引文耦合 365

引文索引 365, 372

引文索引举要 368

引用次序 217

引用信息 200

引证关系排序法 349

引证关系追溯法 82

引证排序法 349

隐藏标签 118

隐性主题 289

印刷型文献 412

应用本体 51, 92

应用关系 289

《英国的海岸线有多长?》 52

英国国家书目 359

《英国牛津大学图书馆目录》

18

英汉对照索引 273

英汉译名对照表 279

《英美编目条例(第二版)》 132

营业性书目 354

影响关系 289

影响因子 435

映射关系 121

《永乐大典》 16

用 265

用户检索 382

用户界面 32, 458

用户友好性 439

有序化理论 40

语法分析标引法 315

语法信息 48

语句 103

语句分析 307

语句加权 307

语言分析 306

语言复分表 213

语言信息组织 12

语言形式复分表 213

语言学 53

《语言学与情报学》 55

语言转向 47

《语言自述集》 326

语义关系 119

语义检索 93

语义上溯 47

语义网 31,91,98,102

语义网格 36

语义信息 48

语用信息 13

语用信息组织 13
预留空号法 226
预印本 435
元标签集 101
元词 263
元词法 81, 262
《元和姓纂》 17
元建模 112
元数据 5, 26
元数据参考信息 200
元数据对象描述模式 180
元数据输入与传输标准 33
元数据著录法 124
元搜索引擎 380
元素限定词 175
原始记录 7
原文来源文献 365
《原子能科技资料主题词典》
25
约束属性 108
月相纪日法 346
韵部排检法 325
韵书 325

Z

杂文和一般论文集索引 374
在版编目 23
在线检索词库 394
增词标引 294
《展开式分类法》 19
站外资源 435
张琪玉 55, 78
张素芳 52
章学诚 17
《哲学词汇》 50
哲学社会科学 68
整体标引 286
整序 4
整序功能 80
正式主题词叙词 272
郑樵 16

政府报告和新闻信息全文数据库 358
政书 412
之嘉 21
支持向量机 304
知识 2
知识本体 90
知识分类法 13
知识管理论 50
知识基础论 48
知识集合论 49
知识交流论 49
知识交流学说 49
知识节 394
知识论 47
知识整体论 47
知识组织 33, 49
《知识组织》 49
知识组织大会 34
知识组织系统 75
直接索引 216
直接转换 290
值域 107
职号 79, 292
职能符号 263
职能符号 78
智能化信息检索 318
智能检索 30
智能搜索引擎 381
智能信息检索 25
《中国大百科全书·哲学卷》 50
《中国编目规则(修订版)》 138
《中国标准文献分类法》 24
《中国大百科全书(简明版)》
426
《中国大百科全书》 67, 419
《中国档案分类法》 24
《中国法》 209, 213, 280, 284
《中国分类主题词表》 25, 280,
284
中国高等教育文献保障系统

358
中国机读目录 143
《中国机读目录格式》 143
《中国警察科学引文索引》 366
《中国科学计量指标: 期刊引证
报告》 368
《中国科学引文数据库》 366
中国科学院国家科学数字图书
馆 407
《中国科学院图书馆图书分类
法》 24
《中国农业科学叙词表》 96
《中国人民大学图书馆图书分
类法》 24
《中国十进分类法及索引》 20
中国数据图书馆工程建设联席
会议 33
《中国图书馆分类法》 20, 24,
64, 214, 240
《中国文献编目规则》 124, 138
中国知网 393
中国字庋擷法 331
《中华人民共和国学科分类与
代码国家标准》 65
中华图书馆协会 21
《中经新簿》 16
《中图法》 24, 240
《中文标题总录初稿》 21
《中文科技期刊数据库》 395
中文目录举要 356
《中文社会科学引文索引》 366
中文索引列举 367
《中文图文标题法》 24
终止符 149
钟义信 47
重点标引 287
重心法 302
周文俊 49
周晓英 45
逐词排列法 332
逐字母排列法 332

- 主表 272
主题 101, 259
《主题编目手册: 标题表》 267
主题标引 286, 293, 297
主题标引法 246
主题词 259, 264, 273, 282
主题词标引 317
主题词表 96
主题词串 282
《主题词-分类号对应表》 280
主题词化 278
主题词排检法 338
主题词索引 216
主题词组配标题 283
主题法 28, 259, 297
主题法系统 54
主题分析 288
主题概念 288, 293
主题结构 290
主题树方法 11
主题索引 366
主题图 28
主题图语言 101
主题图组织 101
主题网关 406
主题语言 54, 80
主题自动标引 86
主题组织法 10
主体 102
主体因素 290
主要款项 135
主要信息源 126
主要主题 289
主页 436
属性 103, 107, 188
属性特性 114
属性约束 113
属于框架 118
助记性 220
注释 118, 214, 268
注释属性 119
注销出版收录情况 387
注音字母排检法 324
注音字母音序法 324
著录 124
著录单元说明 130
著录法 141
著录格式 137
著录规则 127
著录项目 127, 136, 140
著录信息源 126
著录用标识符 127
著者标引 6
著者索引 366
著者原则 135
专类复分表 213
专利文献 258
专门型搜索引擎 379
专题书目 354
专业化信息检索 318
专业型搜索引擎 379
专业主题 289
专用复分表 262
专指词 261
准确匹配 121
桌面搜索引擎 380
资料 4
《资料法》 286
资源 102, 107, 188
资源加工质量 387
资源描述框架 188
资源描述框架模型 98, 105
资源组织 205
资助项目 435
子类 107
子类关系 108
子属性关系 108
字典 412
字典式目录规则 77
《字典式目录条例》 20
字母标记法 227
字母顺序排检法 332
字顺表 272, 279
字顺法 322
字顺排检法 322
字顺组织法 12
字序法 322
字序排检法 322
自底向上法 95
自顶向下法 95
自动标引 25, 85, 306
自动抽词标引 86, 307
自动分词 299
自动分类 26, 300
自动赋词标引 307
自动赋号标引 86
自动聚类 86, 301
自然语言 32, 81
自然语言标引 85
自然语言处理 85
自然语言检索 87
自相似原则 52
自信心定律 341
自引 366
自由标引 85
自由词标引 294
自由浮动复分 269
《自由浮动复分标题字顺索引》
267
自由文本方法 10
自组织理论 42
自组织映射 28
综合标引 287
综合法 95
综合型搜索引擎 379
综合性书目 354
总论复分表 213
族首词 265
族性检索 10
组配标引 293
组配次序 217
组配因素 283

二、外文索引

- A Proposed Format for a
standardized Machine-
Readable Catalog Record
141
- A.Sigel 49
- AACR 2 132
- Academic Universe 393
- Additional Resources 435
- AGROVOC 96
- AIFIA 44
- Aitchison, J. 25
- alt Lable 118
- Alternative 103
- Alternative 192
- Alternative rules 135
- American Memory 458
- Anglo-American Cataloging
Rules 2nded 132
- AOD 96
- application ontology 92
- Appropriate values 174
- Arlene Gtaylor 52
- Art Index 374
- artificial languages 74
- Arts & Humanities Citation
Index 432
- Asilomar Institute For
Information Architecture 44
- Association 101
- association relationship 100
- Austin, D. 78
- Automatic Indexing 306
- Axiom 92
- Bag 103, 192
- BC2 24
- Bertalanffy, L.Von. 40
- Bibliographical description 124
- Biography Index 374
- bibliography 354
- BioOntology Core Group 100
- BIP 359
- BlackWell 电子期刊 393
- Bliss, H.E. 49
- BNB 360
- Boltzmann 41
- Book Review Digest Plus 375
- Book Review Index 364
- Book Trade Bibliography 355
- Books in Print 359
- Borko, H. 26
- Borst 91
- bottom-up 95
- British National Bibliography
360
- broad Match 121
- Brookers, B.C. 49
- C. N. Mooers 4
- C-Gesner 17
- CA 22, 81
- CALIS 27, 358
- CALIS Current Contents Of
Western Journal 376
- CALIS 西文期刊目次数据库
376
- CALIS 重点学科网络资源导航
门户 407
- Callimachus 17
- Cardinality Restriction 113
- CC 237
- CCC 376
- CCF 142
- CDWA 194
- change Note 119
- Chemical Abstracts 81
- Chemical Titles 298
- China Academic
Library&Information System
358
- China Machine-Readable
Catalogue 143
- China MARC Format 143
- Chinese Library Classification
240
- Chinese National Science
Digital Library 407
- Chinese Science Citation
Database 368
- Chinese Social Science Citation
Index 368
- CIP 23
- Citation 365
- Citation Index 365
- Citation Information 200
- citation order 77
- Citation Rate 435
- Cited Author 365
- Cited papers 365
- Cited Reference 366
- Citing Author 365
- Citing papers 365
- CKML 102
- Class 107
- Class 92
- Clausius, Rudolf 41
- CLC 240
- CNKI 393
- CNMARC 143
- Coates, E.J. 77
- Collection 103
- Colon Classification 207
- COM 356
- combination order 77
- Common Communication
Format 143
- Complexity 110
- Comprehensive Bibliography
355
- computation completeness 12
- Computer Aided Indexing 306
- Computer Output Microfilm
356
- Concept 110, 117
- Conceptual Knowledge Markup

- Language 102
Conceptualization 91
Conlon Classification 237
Constraint Property 108
Contact Information 200
Container 103
Content Standards for Digital
 Geospatial Metadata 198
Control fields 147
Controlled languages 74
Controlled Vocabulary 116
Cover 304
CSA Illumina 376
CSA, Cambridge Scientific
 Abstracts 375
CSCD 366
CSDGM 198
CSDL 407
CSSCI 366
Cutter, Charles A. 77
CYBERDEWEY 27
CycL 94
DAML 94, 110
DAML+OIL 111
DAML-ONT 111
DARPA Agent Markup
 Language 110
DARPA 标记语言 110
Data fields 148
Data Quality Information 199
Datatype Property 113
David Filo 382
DC 26, 100, 171
DCMI 171
DDC 19, 228
decidability 112
Defense Documentation Center
 26
definition 119
Description Logic 110
Descriptive Bibliography 354
descriptor 81
Dewey Decimal Classification
 228
DFW 228
digital landscape 44
Directory 146
disjoint With 112
Distribution Information 199
Document Type Definition 99
Domain 107
domain ontology 92
DTD 99, 159
Dublin Core 100
Dublin Core Metadata Initiative
 171
Dumb-Down Principle 173
EB 421
EBSCOhost 391
ED 228
editorial Note 119
《EI 叙词表》 82
Electronic Dewey 228
Element Refinements 175
Elsevier Science 392
Encoding Scheme 175
Encyclopedia Britannica 421
End Note 432
energy 207, 239
Enterprise 法 95
Entity and Attribute Information
 199
Enumerative Bibliography 354
epistemology 47
equivalent Class 112
ERIC 82
e-Science 36
Essay and General Literature
 Index 374
Essential Science Indicators 432
Evaluative bibliography 354
example 119
Extendibility 94
Extensibility 173
Extensible Markup Language 99
facet formula 239
FGDC 198
formal 91
frame-based 100
FRBR 158
free-text search 84
full-text database 385
Function 92
Functional Property 114
Functional Requirements for
 Bibliographic Records 158
Functionality 110
Funding 435
Garfield, Eugene 84
GB 2901—82 143
GB 3792.1—83 136
GB 3792.2—85 136
GB 3792.3—85 136
GB 3792.4—85 136
GB 6447—86 376
GB/T 13745—92 65
GB3792.1—83 124
GB3792.5—85 136
GB3792.6—86 136
GB3792.7—87 136
generic ontology 92
Google 382, 450
Gorman, Michael 132
Grid Computation 36
Gruber 51
Gruber, Tom 91
Haken, H. 42
Haken, Hermann 43
Hart 304
Hausdorff 53
hid Lable 118
Highly Cited.com 435
history Note 119
home page 436
HTML 98
HTML4.01 99
Hulme, E.W. 210
Humanities Index 373
Hyper Text Marku Language 98
IA 43
Identification Information 199
If appropriate 135

- If necessary 135
IFF 102
IFLA 142
《IFLA 书目录的功能要求》 33
IG 406
Impact Factor 435
in Schema 118
Individuals 110
International Classification 49
Infomine 408
Information Gateway 406
Information Architecture 29, 43
Information Exchange Forum on
Spatial Data 198
Information Flow Framework
102
Information Language 55
Instance 92, 107
Internation scope 174
International Standard
Bibliographic Description 128
International Standard
Classification of Education 67
Intrinsicity 173
Intute 407
Inverse Functional Property 114
ISBD (A) 129
ISBD (CF) 129
ISBD (CM) 129
ISBD (CP) 129
ISBD (ER) 129
ISBD (M) 129
ISBD (NBM) 129
ISBD (PM) 129
ISBD (S) 129
ISBD 124, 134
ISBD(G) 129
ISBD(M) 128
ISBDs 23
ISCED 67
ISKO 33
ISO 134
ISO 2709 142
ISO/IEC FCD 13250:1999 101
ISOTC46 33
ISO 国际标准草案 33
ISSN 134
ISSS 34
J.D.Anderson 49
JCR 432
Journal Citation Reports 432
KACTUS 95
Kaiser, J. 77
keyword in content index 298
keyword out of content index
298
Knowledge Organization 49
Knowledge Organization
Systems 75
Koch 52
KOS 75
KWIC 298
KWOC 298
lable Related 121
Lable Relation 121
Language of Indexing and
Classification ——a
Linguistic Study of Structures
and Functions 55
Larry Page 383
LCC 19, 210, 234
LCSH 81, 267
Leader 144
level 239
LEXIS-NEXIS 学术大全数据
库 393
Library of Congress
Classification 234
Library of Congress Online
Catalog 362
Library of Congress Subject
Headings 81, 267
Linguistic 53
Linguistic and Information
Science 55
Link 112, 178, 350
Logic Formalism 110
Luhn 22
M-Taube 21
Machine Readable Catalogue
141
Mandelbort, B. B. 52
MARC 21, 141, 157
MARC XML 158, 165
MARC XML Schema 159
MARC 著录法 124
Markup Language 97
Match 121
matter 207, 239
member 121
member List 121
Metadata 26, 169
Metadata Object Description
Schema 180
Metadata Reference Information
200
meta-modeling 112
METHONTOLOGY 法 95
Micheal Grunibger 95
Mike Uschold 95
Mills, L. J. 24
Minimal ontology commitment
94
Modeling Primitives 100
Modifiability 173
MODS 180
My Library 32
My Yahoo 32
Naïve Bayes 分类法 305
Name Space 100
narrow Match 121
narrower 119
narrower Transtive 119
NASA 26
National Bibliography 354
National Knowledge Internet
393
National Union Catalog 361
natural language 83
Natural Language Processing 85
Net Resource 112
New China MARC Format

- Manual 143
Newspaper Index 364
NKOS 97, 102, 108, 117, 122
NLP 85
note 119, 350
Nothing 112
NS 100
NUC 361
OAI-PMH 400
OAI 协议 400
Object 102
Object Level 110
Object Property 113
Occurrence 101
OCKB 100
OCLC 23
OCLC 世界书目 362
OIL 94, 110
OML 102
One Click Is Enough 29
One-to-One Principle 174
Online Pubic Access Catalog 356
OntoEdit 94
Ontoknowledge 110
Ontoligua 94
Ontology 29, 50, 90
Ontology Container 110
Ontology Definition 110
Ontology Inference Layer 110
Ontology Interchange Language 110
Ontology Markup Language 102
OPAC 26, 356
Open Archives Initiative
 Protocol for Metadata
 Harversting 400
Open Knowledge Base
 Connectivity 100
Open Uniform Resource
 Locators 400
Open URL 400
Optional addition 135
Optionally 135, 173
OWL 94, 111
OWL DL 111
OWL Full 111
OWL Lite 111
OWL 推理 115
Perez 92
Periodicals Index 363
personality 207, 239
PICS 193
Pinakes 17
Platform for Content Selection 193
Popper, Karl 48
post-controlled vocabulary 88
post-coordination 82
pre Lable 118
PRECIS 24, 78
pre-controlled vocabulary 88
pre-coordination 82
Predicate 102
Preprints 435
PREservwed Context Indexing
 System 78
Prigogine, I. 42
Pro Cite 432
Proceeding Index 365
Properties 188
Property 103, 107, 113
ProQuest 全文数据库 390
Protégé 94
PTLA 359
Publishers Trade List Annual 359
R.S.Wurman 15
Ranganathan, S. R. 78, 207
Range 107
RDF 104, 180, 189
RDF Schema 112
RDF(S) 109, 116, 122
RDF/XML 104
RDFC 98
RDFS 105, 111
Readers' Guide to Periodical
 Literature 373
record terminator 149
Reference Manager 432
related 119
related Match 121
Related Records 366, 434
Relation 92
Renardus 34
Repeatability 173
Research Activities 435
Resource 103, 188
Resource Description
 Framework Schema 98, 105
Role 110
Root Elements 181
round 239
Satisfiability 110
SC9 33
Schrodinger, E. 42
SCI 82
Science Citation Index 82
Science Citation Index
 Expanded 432
scope Note 119
Search Engine 445
Search/Retrieval via URL 181
Semantic Grid 36
semantic Relation 119
Semantic Web 98, 102, 111
Sequence 103, 192
Sergey Brin 383
Service On Demand 29
SGML 98
Shannon 41
shared 91
Shel Kimen 45
Shiri, Ali 36
Sierpinski 52
SIG, Subject Information
 Gateway 406
Simple Knowledge Organization
 System 116
Simplicity of creation and
 maintenance 173

- Skeletal Methodology 95
SKOS 98, 116, 122
Social Science Citation Index 432
Social Science Index 373
SOM 28
Source items 365
SP 406
space 207, 239
Spatial Data Organization Information 199
Spatial Reference Information 199
Specification of elements 130
Springer Link 392
SRI's Artificial Intelligence Centre 100
SRI 研究所 100
SRU 181
Standard Generalized Markup Language 98
Statement 103, 108, 188
Stop-term 82
Studer 91
subClassOf 107
Subject 102
Subject Bibliography 355
Subject Guide to Books in Print 359
subject heading 81
Subject Portal 406
subPropertyOf 108
Subsumption 110
Support Vector Neighbours 304
SVM 304
Symmetri Property 114
Syntax independence 173
task ontology 92
term 100
THCI 366
The Categories for the Description of Works of Art 194
the Federal Geographic Data Committee 198
The International Federation of Library Associations and Institutions 142
The National Centre for Biomedical Ontology 100
The Networked Knowledge Organization System/Service Working Group 100
Thesaurus 82
Thing 112
Tim Berners-Lee 91, 99
time 207, 239
Time Cited 366
Time Period Information 200
Top Concept 118
Top Level Elements 181
top-down 95
Topic 101
Topic Maps 101
TopicMaps.Org 101
TOVE 法 95
TSSCI 366
UDC 232
Ulrich's Periodicals Directory 361
Ulrich 期刊指南 361
UNESCO 142
UNIMARC 142
Union List of Serials in Library of the United States and Canada 362
unit term 81
Universal Decimal Classification 232
Universal MARC Format 142
USE 参照相 269
variable fields 147
Vector Space Model 304
Vickery, B.C. 78
Vocabulary Mark-up Language 100
Voc-ML 100
VSM 304
W3C 34, 98
Wade, Thomas F. 326
Web 2.0 34
Web directory 441
Web of Knowledge 432
Web of Science 369, 372, 432
Web Ontology 111
Web Ontology Language 94, 111
web pages 436
Web site Reviews 435
Website 112
Wiley Inter Science 电子期刊 393
Wilson Select 393
Wilson 公司索引数据库 373
Winkler, Paul W. 132
WOK 432
Wolf, Chrisian 50
WordNet 52
World Cat 362
World Wide Consortium 99
World Wide Web Consortium 98
Wurman, R. S. 43
WWW. Virtual Library 27
XML 99, 108, 178, 189
XML Schema 162
XML/RDF(S) 120
XML-Based Ontology Exchange Language 100
XOL 100
XTM 101
Yahoo 382, 444
Z39.19 122
Z39.50 33, 399
Ziga Turk 36
Zipf 定律 306


三、数字索引

0 251

1:1 原则 174

《2005 年世界主要语种、分布
和应用力调查报告》 322

748 工程 25



参 考 文 献

- [1] 白献阳, 田立忠. 学科信息门户评价指标体系研究[J]. 情报理论与实践, 2006 (1): 37~39, 43.
- [2] 北京蚁巢科技有限责任公司. 中国大百科全书简明版光盘版. 北京: 中国大百科全书出版社, 2004.
- [3] 曹建. Internet 与 E-mail 安全防范实用技术[M]. 成都: 电子科技大学出版社, 1999.
- [4] 曹伟. 信息环境论的产生[J]. 东岳论丛, 2000, 21 (5): 84~86.
- [5] 陈楚祥. 词典评价标准十题[J]. 辞书研究, 1994 (1): 10~21.
- [6] 陈嘉明. 当代知识论: 概念、背景与现状[J]. 哲学研究, 2003 (05): 89~95.
- [7] 陈金莉, 等. 数字图书馆标准与规范建设: Z39.50 协议应用指南[R]. 2004.
- [8] 陈朋. ISI Web of Knowledge 集成检索平台探析[J]. 图书馆杂志, 2004 (9): 55~59.
- [9] 陈桃, 明均仁. 现代信息检索技术发展趋势初探[J]. 农业图书情报学刊, 2007, 19 (6).
- [10] 陈雅芝. 信息检索[M]. 北京: 清华大学出版社, 2006.
- [11] 陈耀盛. 网络信息组织[M]. 北京: 科学技术文献出版社, 2004.
- [12] 程小澜, 泮杏梅. 光盘数据库的情报价值与评价选择[J]. 现代图书情报技术, 1998 (4): 34~37.
- [13] 储节旺, 郭春侠, 吴昌合. 信息组织学[M]. 北京: 清华大学出版社, 2007.
- [14] 储节旺, 郭春侠. 信息组织: 原理、方法和技术[M]. 合肥: 安徽大学出版社, 2002.
- [15] 戴维民, 田春光. 从情报检索语言到本体——信息组织的新变革[J]. 图书情报工作, 2005 (07): 6~10.
- [16] 戴维民. 信息组织[M]. 北京: 高等教育出版社, 2004.
- [17] 戴维民, 等. 文献信息数据库建库技术[M]. 北京: 北京图书馆出版社, 2001.
- [18] 丁璇, 侯汉清, 章成志. 中文网页标引源主题表达能力的调查统计[J]. 大学图书馆学报, 2002, (6).
- [19] 董慧, 丁波涛. OAI-MHP 协议初探[J]. 图书情报知识, 2004 (6): 70~73.
- [20] 段明莲, 沈正华. 数字时代的图书馆信息资源组织[M]. 北京: 北京图书馆出版社, 2006.
- [21] 费玉梅. 语法现象分析在汉语文献主题标引中的作用[J]. 无锡教育学院学报. 2003, 23 (9).
- [22] 冯惠玲, 王立清. 信息检索教程[M]. 中国人民大学出版社, 2004.
- [23] 傅守灿, 陈文广. 图书馆自动化基础教程[M]. 北京: 北京大学出版社, 1996.
- [24] F. W. Lancaster. 情报检索词汇控制 (第 2 版) [M]. 侯汉清, 等, 译. 上海: 同济大学出版社, 1992.
- [25] F. W. Lancaster. 自然语言检索和后控词表[J]. 桑仁义, 译. 图书情报通讯, 1991 (2): 45~48, 37.
- [26] 高红, 顾桦. MARC21 规范数据格式使用手册[M]. 北京: 北京图书馆出版社, 2005.
- [27] 耿海英, 肖仙桃. Web of Science 和 Google Scholar 引文检索功能比较[J]. 图书与情报, 2007 (3): 100~102.
- [28] 龚蛟腾. 传统文献组织与网络信息组织之比较[J]. 四川图书馆学报, 2004 (1): 21~24.
- [29] 龚立群, 孙洁丽. OAI、SRW/U 及 OpenURL 的比较及协同使用研究[J]. 情报科学, 2007, 25 (7): 1073~1079.
- [30] 国家图书馆《中国图书馆分类法》编辑委员会. 中国图书馆分类法 (第 5 版) [M]. 北京: 国家图书馆出版社, 2012.
- [31] 国际标准化组织 (ISO). 文献与情报工作国际标准汇编[M]. 北京: 科学技术文献出版社, 1983: 7.
- [32] 韩客松, 王永成. 一种用于主题提取的非线性加权方法[J]. 情报学报, 2000, (6).
- [33] 韩客松, 王永成. 中文全文标引的主题词标引和主题概念标引方法. 情报学报, 2001, (4).

- [34] 韩立栋. 基于 XML 的 MARC 探讨[J]. 兰台世界, 2005 (13): 60~61.
- [35] 何嘉, 陈琳. 基于神经网络汉语分词模型的优化[J]. 成都信息工程学院学报, 2006, 21 (12).
- [36] 侯汉清, 章成志, 郑红. Web 概念挖掘中标引源加权方案初探[J]. 情报学报, 2005, (1).
- [37] 侯汉清. 索引技术和索引标准[M]. 北京: 北京图书馆出版社, 1997.
- [38] 胡昌平, 等. 面向用户的信息资源整合与服务[M]. 武汉: 武汉大学出版社, 2007.
- [39] 胡昌平, 等. 信息资源管理研究进展[M]. 武汉: 武汉大学出版社, 2008.
- [40] 胡双演, 李俊山, 李建军. 基于潜在语义分析的视频检索[M]. 计算机工程, 2007 (13).
- [41] 华薇娜. 网络学术信息资源检索与利用[M]. 北京: 国防工业出版社, 2002.
- [42] 黄建年. MARC 数据与图书馆[J]. 津图学刊, 1997 (4): 26~32.
- [43] 黄如花, 刘鑒. WorldCat 标注系统的优化方案[J]. 图书与情报, 2012 (5): 16~20.
- [44] 黄如花. 数字图书馆信息资源的建设与管理[M]. 武汉: 武汉大学出版社, 2005.
- [45] 黄如花. 数字图书馆信息组织的评价[J]. 情报杂志, 2005 (7): 7~9.
- [46] 黄如花. 网络信息的检索与利用[M]. 武汉: 武汉大学出版社, 2002.
- [47] 黄如花. 网络信息组织: 模式与评价[M]. 北京: 北京图书馆出版社, 2003.
- [48] 黄如花. 网络信息组织的发展趋势[J]. 中国图书馆学报, 2003 (4).
- [49] 黄如花. 学科信息门户信息组织的评价[J]. 武汉大学学报 (社会科学版), 2003 (9): 653~657.
- [50] 黄如花. 学科信息门户信息组织的优化[J]. 图书情报工作, 2005 (7): 11~15, 10.
- [51] 黄水清, 程冲, 李志燕. 开放式非相关文献知识发现方法在中文文献中的验证[J]. 情报理论与实践, 2008, 31 (2).
- [52] 冀颖. 论自然语言检索系统中的控制问题[J]. 晋图学刊, 2006 (1): 14~17.
- [53] 贾晓斌, 魏杰. 《中国大百科全书》特征述评[J]. 图书馆论坛, 1996 (6): 10~12.
- [54] 蒋永福. 图书馆与知识组织——从知识组织的角度理解图书馆[J]. 中国图书馆学报, 1999 (05): 19~23.
- [55] 焦玉英. 信息检索进展[M]. 北京: 科学出版社, 2003.
- [56] 金常政. 论百科全书的评价标准[J]. 辞书研究, 1990 (1): 5~13.
- [57] 靳从, 唐振民, 杨静宇. 自动标引中自然主题词的切分. 情报科学, 2004, 22, (3).
- [58] 阚华, 高路克, 周红雁. 信息组织[M]. 合肥: 安徽科学技术出版社, 2007.
- [59] 柯平, 高洁. 信息管理概论[M]. 上海: 科学出版社, 2007: 5~6.
- [60] [西德]克劳斯·赖本齐. 实用情报文献工作原理[M]. 来新枚, 等译. 北京: 科学技术文献出版社, 1983.
- [61] 赖茂生, 王延飞, 赵丹群. 计算机情报检索[M]. 北京: 北京大学出版社, 2001.
- [62] 冷伏海. 信息组织概论[M]. 北京: 科学出版社, 2003.
- [63] 李宏伟, 等. 网络地理信息系统与空间元数据[M]. 郑州: 黄河水利出版社, 2004.
- [64] 李后卿, 柳晓春. 图书情报学理论基础中的知识论研究[J]. 中国图书馆学报, 2003 (01): 82~84.
- [65] 李伟, 黄颖. 文本聚类算法的比较[J]. 科技情报开发与经济, 2006, 16 (22).
- [66] 李希明, 土丽艳, 金科. 从信息孤岛的形成谈数字资源整合的作用[J]. 图书馆论坛, 2003, 23 (6): 121~122, 66.
- [67] 李晓明, 闫宏飞, 王继民. 搜索引擎——原理、技术与系统[M]. 北京: 科学出版社, 2005.
- [68] 李新华. 各种后控词表实现模式的分析[J]. 图书馆杂志, 2003, 22 (4): 12~15.
- [69] 李醒民. 论科学的分类[J]. 武汉理工大学学报 (社会科学版), 2008. (2).
- [70] 酈金花. 基于 CNMARC 字典库和 XML 的 MARC 发布系统的设计[J]. 情报杂志, 2006 (12): 87~89.
- [71] 刘炜. DC 元数据年度进展[J]. 数字图书馆论坛, 2006 (11).
- [72] 刘刚. 单汉字标引检索研究述评[J]. 图书馆建设, 1999 (6).
- [73] 刘海峰, 王倩, 王元元. 基于 Web 的文本检索位置加权模型研究[J]. 情报学报, 2007, 25 (3).
- [74] 刘嘉. 网络信息资源的组织: 从信息组织到知识[M]. 北京: 组织北京图书馆出版社, 2002.

- [75] 刘建平. 医学网络实用技术教程[M]. 北京: 中国铁道出版社, 2007.
- [76] 刘荣. 图书情报管理自动化基础[M]. 武汉: 武汉大学出版社, 1998.
- [77] 刘务华, 罗铁坚, 王文杰. 文本聚类算法的质量评价[J]. 中国科学院研究生院学报, 2006, 23 (5).
- [78] 马费成. 导言: 情报学中的序[J]. 图书·情报·知识, 2008 (3).
- [79] 马费成等. 信息管理学基础[M]. 武汉: 武汉大学出版社, 2002.
- [80] 马然, 侯汉清. 基于引文的自动标引法初探[J]. 江苏图书馆学报, 2002, (1).
- [81] 马张华, 黄智生. 网络信息资源组织[M]. 北京: 北京大学出版社, 2007.
- [82] 马张华. 信息组织 (第3版) [M]. 北京: 清华大学出版社, 2008.
- [83] 茆意宏, 张俊, 黄水清. 异构数据库互操作协议: Z39.50 和 OpenURL 的比较研究[J]. 图书馆理论与实践, 2006 (4): 101~104.
- [84] 美国不列颠百科全书公司. 不列颠百科全书国际中文版[K]. 北京: 中国大百科全书出版社, 2007.
- [85] 聂玮. 谈《现代汉语词典》第5版相关体例问题[J]. 湖北社会科学, 2008 (3): 139~141.
- [86] 潘有能. 一个自动分词分类系统的实现[J]. 情报学报, 2002, 21 (2).
- [87] 彭冬莲. 单汉字标引及其检索技术的优化[J]. 农业图书情报学刊, 2005, 17 (4).
- [88] 彭斐章. 目录学教程[M]. 北京: 高等教育出版社, 2004.
- [89] 邵献图. 西文工具书 (第3版). 北京: 北京大学出版社, 1998.
- [90] 沈固朝. 网络信息检索: 工具、方法、实践[M]. 北京: 高等教育出版社, 2004.
- [91] 沈固朝. 信息检索 (多媒体) 教程[M]. 北京: 高等教育出版社, 2002.
- [92] 沈芸芸, 裴微微. 国家图书馆核心元数据标准著录规则 (初稿) [R]. 国家图书馆, 2009, 9.
- [93] 沈正华. 光盘数据库质量的评估[J]. 现代图书情报技术, 1996, (4): 38~44.
- [94] 施术才. 自动标引的研究现状与和的发展方向[J]. 中国计算机用户, 1990, (12).
- [95] 史继红, 赖茂生. 汉语自动标引加权方法试验研究[J]. 现代图书情报技术, 1994, (3).
- [96] 史田华等. 信息组织与存储[M]. 南京: 东南大学出版社, 2003.
- [97] 史文中. 空间数据与空间分析不确定性原理[M]. 北京: 科学出版社, 2005.
- [98] 宋炜, 张铭. 语义网简明教程[M]. 北京: 高等教育出版社, 2004.
- [99] 苏新宁, 邵波. 信息传播技术[M]. 南京: 南京大学出版社, 1998.
- [100] 苏新宁, 邹晓明. 文献信息自动标引研究[J]. 数字图书馆技术, 2000 (1).
- [101] 苏新宁. 视频信息索引技术研究进展[J]. 情报学报, 2004, 23 (8).
- [102] 孙更新. 文献信息编目[M]. 武汉: 武汉大学出版社, 2006: 454.
- [103] 孙一中. XML 理论和应用基础[M]. 北京: 北京邮电大学出版社, 2000.
- [104] 唐琼, 张玫. “美国记忆”与“共享工程”比较研究[J]. 图书馆理论与实践, 2006 (1): 7~9.
- [105] 唐炜. 大型综合性中文门户网站信息组织体系分析[J]. 图书情报工作, 2005 (2): 27~31.
- [106] 田苗苗, 许建, 汪津, 丁桂英. 基于遗传算法的 Web 信息自动标引研究[M]. 吉林大学学报. 2006, 24 (5).
- [107] 涂子沛. 大数据: 正在到来的数据革命, 以及它如何改变政府、商业与我们的生活[J]. 桂林: 广西师范大学出版社, 2012: 35.
- [108] [美]泰勒. 信息组织[M]. 张素芳, 等, 译. 北京: 机械工业出版社, 2006.
- [109] 汪媛, 赖茂生. 我国高校图书馆引进网络版全文数据库的综合评价模型[J]. 情报科学, 2004, 22 (9): 1061~1065.
- [110] 王剑波. 论 MARC 编目网络信息资源——必要性、可行性及实践[J]. 图书与情报, 2006, (2).
- [111] 王军, 张丽. 网络知识组织系统的研究现状和发展趋势[J]. 中国图书馆学报, 2008 (1): 65~69.
- [112] 王兰成, 蒋丹, 刘庆辉. 全文数据库建库原理与应用技术[J]. 情报学报, 1999, 18 (4): 321~328.
- [113] 王荣江. 奎因的整体知识论及其后果[J]. 自然辩证法研究, 2007 (02): 40~43.
- [114] 王身立. 耗散结构理论向何处去——广义进化与负熵[M]. 北京: 人民出版社, 1989.
- [115] 王卫军. Web2. 0 环境下网络信息的组织[J]. 中国科技资源导刊, 2009 (4): 47~51.
- [116] 王子舟. 知识集合初论——对图书馆学研究对象的探索[J]. 中国图书馆学报, 2000 (04): 7~12.

- [117] 文庭孝. 汉语自动分词研究进展[J]. 图书情报, 2005, (5).
- [118] 文庭孝. 情报检索中汉语语词自动切分研究[J]. 图书与情报, 2001, (2).
- [119] 吴翠兰. 当前图书馆编目工作的发展契机. <http://www.chnlib.com/Zylwj/2709.html> (2008-12-4)
- [120] 吴桂英. MODS 的构成、特点及其应用前景探索[J]. 中小学图书情报世界, 2008 (1): 17~19.
- [121] 吴建中, 侯汉清. 从人工语言到自然语言——关于图书馆未来的对话之十[J]. 图书馆杂志, 1996 (4): 32~34.
- [122] 吴建中. DC 元数据[M]. 上海: 上海科学技术文献出版社, 2000.
- [123] 吴启明, 易云飞. 文本聚类综述[J]. 河池学院学报, 2008, 28 (2).
- [124] 吴万晔. 论 MARC 元数据的缺陷及发展趋势[J]. 图书馆工作与研究, 2006 (2): 28~29.
- [125] 伍玉伟. 信息构建 (IA) 与信息组织的比较研究[J]. 图书馆论坛, 2006 (04): 49~51.
- [126] 肖炳珠. 基于视频内容的检索方法[J]. 现代电视技术, 2001, (12).
- [127] 肖珑. 互联网上的全文数据库与全文服务[J]. 大学图书馆学报, 2000 (3): 3~8.
- [128] 肖珑. 美国国家数字图书馆项目的进展[J]. 情报学报, 1998 (3): 190~196.
- [129] 肖希明等. 数字信息资源建设与服务研究[M]. 武汉: 武汉大学出版社, 2008.
- [130] 谢琴芳. CALIS 联机合作编目手册[M]. 北京: 北京大学出版社, 2000.
- [131] 熊回香, 金晓耕. Web2.0 环境下信息组织的优化研究——以豆瓣网为例[J]. 现代情报, 2012 (4): 19~24.
- [132] 熊回香, 夏立新. 汉语分词技术综述[J]. 图书情报工作, 2008, 52, (4).
- [133] 熊回香. Web2.0 环境下的网络信息组织[J]. 情报资料工作, 2007 (5): 29~32, 50.
- [134] 徐秉静, 詹剑, 贺前华. 基于神经网络的分词方法[J]. 中文信息学报, 1993 (2).
- [135] 徐天秀. 信息检索[M]. 北京: 科学出版社, 2006.
- [136] 延卫平. MARC 的 XML 交换格式研究[J]. 现代图书情报技术, 2006 (8): 31~35.
- [137] 严红. 百科全书编排方式比较研究[J]. 武汉大学学报, 1996 (3): 116~122.
- [138] 杨芳, 杨振山. 基于语义网技术的主题词自动标引[M]. 计算机工程与设计, 2005, 26, (10).
- [139] 杨嫚. 网络信息资源组织与开发研究[M]. 武汉: 华中科技大学出版社, 2006.
- [140] 杨玉麟. 信息描述[M]. 北京: 高等教育出版社, 2004.
- [141] 叶继元. 信息检索导论[M]. 北京: 电子工业出版社, 2009.
- [142] 叶志清, 刘瑞红, 袁庆, 胡修兰. 文献信息计算机全文全自动标引方法[J]. 情报学报, 2003, 22 (4).
- [143] 应峻, 徐一新. 网络全文数据库资源评价[J]. 现代图书情报技术, 2005 (3): 57~59, 15.
- [144] 于明洁. 豆瓣网 Tag 模式对图书馆信息组织的启示[J]. 数字图书馆论坛, 2009. (12): 99~102.
- [145] 余翠莉, 徐军英. Yahoo 和 Google 搜索功能之比较[J]. 农业图书情报学刊, 2007. (6): 109~111.
- [146] 俞宣孟. 本体论研究[M]. 上海: 上海人民出版社, 2005.
- [147] 岳剑波. 我国信息环境管理的政策调控与信息立法问题[J]. 情报资料工作, 2000 (3): 6~9.
- [148] 岳珍. 四大中文搜索引擎检索性能测评[J]. 情报科学, 2005 (6): 921~927.
- [149] 臧国全. 浅析信息资源数字化的存储技术[J]. 情报科学, 2000, 18 (12): 1113~1115.
- [150] 曾蕾. 网络环境下的知识组织系统——编者的话[J]. 现代图书情报技术, 2004, (1): 2~3.
- [151] 詹德优. 中文工具书导论[M]. 武汉: 湖北教育出版社, 1994.
- [152] 湛垦华. 普利高津与耗散结构理论[M]. 西安: 陕西科学技术出版社, 1998.
- [153] 张冬慧, 孙波, 徐照财, 程显毅. 文本自动分类关键技术研究[J]. 微计算机信息, 2008, 2 (3).
- [154] 张帆. 信息组织学[M]. 北京: 科学出版社, 2005.
- [155] 张帆, 等. 信息存储与检索[M]. 北京: 高等教育出版社, 2003: 440.
- [156] 张会平等. 概念图在知识组织中的应用研究[J]. 情报科学, 2007 (10).
- [157] 张家耕, 谢晓竹. XML 网络编程技术[M]. 北京: 国防工业出版社, 2002.
- [158] 张靖. XML/RDF 在图书馆元数据 MARC 中的应用描述[J]. 微计算机信息, 2008 (15): 193~195.
- [159] 张琪玉, 侯汉清. 情报检索语言实用教程[M]. 武汉: 武汉大学出版社, 2004.
- [160] 张琪玉. 论后控制词表[J]. 图书情报工作, 1994 (1): 1~4.

- [161] 张琪玉. 情报语言学基础[M]. 武汉: 武汉大学出版社, 1997.
- [162] 张琪玉. 情报检索语言[M]. 武汉: 武汉大学出版社, 2004. 7.
- [163] 张晓林. 元数据研究与应用[M]. 北京: 北京图书馆出版社, 2002.
- [164] 张艳梅, 胡文淑, 曾锡. 基于神经网络的中文分词技术研究[J]. 软件导刊, 2007, (12).
- [165] 张燕飞. 信息组织的主题语言[M]. 武汉: 武汉大学出版社, 2005.
- [166] 章成志. 自动标引研究的回顾与展望[J]. 现代图书情报技术, 2007, (11).
- [167] 赵玲. 网上信息资源与光盘数据库之比较[J]. 情报资料工作, 2002 (3): 37~38.
- [168] 赵妍, 侯汉清, 耿金玉, 等. 中文期刊论文自动标引加权设计研究[J]. 新世纪图书馆, 2004, (1).
- [169] 赵岩碧. 信息检索原理与方法教程[M]. 北京: 化学工业出版社, 2005.
- [170] 赵正文, 康耀红. Web 信息检索结构化排序函数与标引词加权技术[J]. 计算机工程与应用, 2007, 43 (11).
- [171] 真湊. 美国记忆特点、技术方案要点及质量标准(上)[J]. 情报理论与实践, 2001 (4): 313~315.
- [172] 真湊. 美国记忆特点、技术方案要点及质量标准(下)[J]. 情报理论与实践, 2001 (5): 393~395.
- [173] 中国 21 世纪议程管理中心. Internet 与可持续发展网络实用教程[M]. 北京: 科学出版社, 1998.
- [174] 中国大百科全书出版社编辑部. 中国大百科全书(简明版)[K]. 北京: 中国大百科全书出版社, 2004.
- [175] 中国社会科学院语言研究所词典编辑室. 现代汉语词典(第 5 版)[K]. 北京: 商务印书馆, 2005.
- [176] 钟义信. “知识论”基础研究[J]. 电子学报, 2001 (01): 96~102.
- [177] 钟义信. 知识论: 核心问题——信息—知识—智能的统一理论[J], 2001 (04): 526~530.
- [178] 周黎明, 张洋. 基于信息环境论的信息环境管理[J]. 图书馆论坛, 2005, 25 (2): 24~28.
- [179] 周宁. 信息组织学教程[M]. 北京: 科学出版社, 2007.
- [180] 周宁. 信息组织[M]. 武汉: 武汉大学出版社, 2001.
- [181] 周晓英. 基于信息理解的信息构建[M]. 北京: 中国人民大学出版社, 2005.
- [182] 朱礼军, 赵新力, 乔晓东, 等. 跨领域多来源主题词表集成与服务研究[J]. 现代图书情报技术, 2007, (1).
- [183] 朱丽. 自然语言的应用研究[J]. 图书与情报, 1996 (3): 34~37.
- [184] 总装备部. 卫星应用现状与发展[M]. 北京: 中国科学技术出版社, 2001.
- [185] [美]Arlene G. Taylor. 信息组织[M]. 张素芳, 等, 译. 北京: 机械工业出版社, 2006.
- [186] 王美琴. 学科信息门户可持续性发展研究[D]. 南京: 南京农业大学, 2007.
- [187] 百度百科. 富媒体[EB/OL]. [2013-8-8]. <http://baike.baidu.com/view/184217.htm>
- [188] 百度百科. 学科信息门户[EB/OL]. [2013-8-8]. <http://baike.baidu.com/view/626891.htm>
- [189] 百度百科[EB/OL]. [2013-08-23]. <http://baike.baidu.com/>.
- [190] 都柏林核心元数据元素集, 1.1 版本. [http://dc.library.sh.cn/dces1999.htm\(2008-12-22\)](http://dc.library.sh.cn/dces1999.htm(2008-12-22))
- [191] 豆瓣网[EB/OL]. [2013-08-22]. <http://www.douban.com/>.
- [192] 各种元数据格式简介. [http://www.lib.hust.edu.cn/lib/dllib.nsf/1ce930115fcd2e7048256c9a0006f537/52809febff2fd74348256c9f002b29d5?OpenDocument\(2009-1-7\)](http://www.lib.hust.edu.cn/lib/dllib.nsf/1ce930115fcd2e7048256c9a0006f537/52809febff2fd74348256c9f002b29d5?OpenDocument(2009-1-7))
- [193] 关于 DC-2005. [http://blog.donews.com/kevenlw/archive/2005/11/11/623610.aspx\(2008-12-25\)](http://blog.donews.com/kevenlw/archive/2005/11/11/623610.aspx(2008-12-25))
- [194] 刘炜, 楼向英, 赵亮. DC 元数据的历史、现状及未来. [http://eprints.rclis.org/archive/00003408/01/DCMI4%E5%B9%B4%E5%88%8A_DC.pdf\(2008-11-17\)](http://eprints.rclis.org/archive/00003408/01/DCMI4%E5%B9%B4%E5%88%8A_DC.pdf(2008-11-17))
- [195] 刘炜. 关于元数据的十万个为什么. [http://www.libnet.sh.cn/sztsg/fulltext/abc/metaFAQ.pdf\(2008-12-7\)](http://www.libnet.sh.cn/sztsg/fulltext/abc/metaFAQ.pdf(2008-12-7))
- [196] 美国信息集团 CSA 数据库系统介绍, available from: <http://www.csa.com>
- [197] MARC21 书目格式介绍. [http://210.32.137.28/calis/download/train/20051010/MARC21%E4%B9%A6%E7%9B%AE%E6%A0%BC%E5%BC%8F%E8%AF%A6%E7%BB%86%E4%BB%8B%E7%BB%8D.doc\(2008-11-17\)](http://210.32.137.28/calis/download/train/20051010/MARC21%E4%B9%A6%E7%9B%AE%E6%A0%BC%E5%BC%8F%E8%AF%A6%E7%BB%86%E4%BB%8B%E7%BB%8D.doc(2008-11-17))
- [198] 潘岩铭, 金培华. 从机读目录格式到通用记录格式: 面向未来的信息服务. [http://www.marsoft.cn/marc/article6.htm\(2008-11-16\)](http://www.marsoft.cn/marc/article6.htm(2008-11-16))

- [199] 维基百科[EB/OL]. [2013-08-23]. <http://zh.wikipedia.org/>.
- [200] 温琳琳. 资讯组织研究. [http://203.208.37.104/translate_c?hl=zh-CN&sl=zh-TW&u=http://river.glis.ntnu.edu.tw/homework/information-1/work/MARC21_P.pdf&prev=/search%3Fq%3DMARC%2B21%2B%25E4%25B9%25A6%25E7%25B9%25AE%2B%25E5%2586%2585%25E5%25AE%25B9%26star%25D20%26hl%3Dzh-CN%26newwindow%3D1%26sa%3DN&usg=ALkJrhiS3EH8n8GV-lwXEwNRDjnZKJgUiA\(2008-12-2\)](http://203.208.37.104/translate_c?hl=zh-CN&sl=zh-TW&u=http://river.glis.ntnu.edu.tw/homework/information-1/work/MARC21_P.pdf&prev=/search%3Fq%3DMARC%2B21%2B%25E4%25B9%25A6%25E7%25B9%25AE%2B%25E5%2586%2585%25E5%25AE%25B9%26star%25D20%26hl%3Dzh-CN%26newwindow%3D1%26sa%3DN&usg=ALkJrhiS3EH8n8GV-lwXEwNRDjnZKJgUiA(2008-12-2))
- [201] 新浪网[EB/OL]. [2013-08-10]. <http://www.sina.com.cn>.
- [202] 元数据. [http://baike.baidu.com/view/107838.htm\(2008-12-7\)](http://baike.baidu.com/view/107838.htm(2008-12-7))
- [203] 在HTML中使用DC元数据. [http://www.chinaitpower.com/A200508/2005-08-02/178448.html\(2008-12-29\)](http://www.chinaitpower.com/A200508/2005-08-02/178448.html(2008-12-29))
- [204] 曾新红, 蔡庆河. OWL2 Web 本体语言快速参考指[EB/OL]. (2012-09-25)[2013-06-10]. <http://nkos.lib.szu.edu.cn/OWL2/OWL2QuickReferenceGuideSimplifiedChinese.htm>
- [205] 张慧蓉. 浅谈 MARC 的发展趋势. [http://www.lib.ntu.edu.tw/pub/mk/mk55/mk55-09.htm\(2008-12-3\)](http://www.lib.ntu.edu.tw/pub/mk/mk55/mk55-09.htm(2008-12-3))
- [206] 《中国图书馆分类法》(第五版) 修订重点[2011-07-05]. <http://wenku.baidu.com/view/6b13bd1d650e52ea55189853.html>.
- [207] 中华人民共和国学科分类与代码简表. (国家标准 GBT 13745-2009). [2012-03-12]. <http://wenku.baidu.com/view/53fb1bee6294dd88d0d26b28.html>.
- [208] 中国科学引文索引数据库介绍, available from: http://sdb.csdli.ac.cn/index_more1.jsp
- [209] 中文社会科学引文索引数据库介绍, available from: <http://cssci.nju.edu.cn/introduce.htm>
- [210] Agrovoc Web Services Version 2.0 Documentation [EB/OL]. [2010-07-28]. <http://aims.fao.org/website/Documentation/sub>.
- [211] American Memory[EB/OL]. [2013-08-16]. <http://memory.loc.gov/>.
- [212] Berners - Lee, Tim. CoolUR Is don't Change. [EB/OL].[2008-06-20]. <http://www.w3.org/Provider/Style/URI>
- [213] Berners-Lee, T. Linked data – design issues [EB/OL]. [2010-07-01]. <http://www.w3.org/DesignIssues/LinkedData.html>.
- [214] Berners-Lee, T.Design Issues: What the Semantic Web can represent [EB/OL]. [2014-03-03]. <http://www.w3.org/DesignIssues/RDFnot.html>.
- [215] Bizer, C., Cyganiak, R., Heath, T. How to Publish Linked Data on the Web [EB/OL]. [2010-07-01]. <http://www4.wiwi.fu-berlin.de/bizer/pub/LinkedDataTutorial/>.
- [216] CALIS 西文期刊目次数据库介绍, available from: <http://ccc.calis.edu.cn>
- [217] D2R Server: Accessing databases with SPARQL and as Linked Data.[EB/OL].[2010-07-01].<http://d2rq.org/d2r-server>
- [218] D'Arcus, B. & Giasson, F. Bibliographic Ontology Specification [EB/OL].[2010-04-28]. <http://bibliontology.com/specification>
- [219] DCMI 元数据术语. [http://dc.library.sh.cn/dcmi-terms.htm#part3\(2008-12-22\)](http://dc.library.sh.cn/dcmi-terms.htm#part3(2008-12-22))
- [220] EUROBroker S.Thesaurus guide: Analytical directory of selected vocabularies for information retrieval, 1992 (2nd version)[R].Luxembourg: European Communities, 1993.
- [221] Frequently-asked questions on FGDC metadata. [http://geology.usgs.gov/tools/metadata/tools/doc/faq.html#q1.1\(2008-12-7\)](http://geology.usgs.gov/tools/metadata/tools/doc/faq.html#q1.1(2008-12-7))
- [222] Google[EB/OL]. [2013-08-15]. <http://www.google.com>.
- [223] Golub K, Tudhope D.Terminology registry scopingstudy(TRSS):Final report[R/OL]. UK: Joint Information Systems Committee(JISC), 2009.[2011-01-28]<http://www.jisc.ac.uk/media/documents/programmes/sharedservices/trss-report-final.pdf>.
- [224] Guidelines for implementing Dublin Core in XML.[http://dublincore.org/documents/2003/04/02/dc-xml-guidelines/\(2008-12-29\)](http://dublincore.org/documents/2003/04/02/dc-xml-guidelines/(2008-12-29))
- [225] Harper, Corey. Authority Control for the Semantic Web. Encoding Library of Congress Subject

- Headings. International Conference on Dublin Core and Metadata Applications [EB/OL], Manzanillo, Mexico.[2008-06-20].<http://hdl.handle.net/1794/3268>.
- [226] Harper, Corey. Authority Control for the Semantic Web. Encoding Library of Congress Subject Headings. International Conference on Dublin Core and Metadata Applications [EB/OL], Manzanillo, Mexico. [2008-06-20]<http://hdl.handle.net/1794/3268>.
- [227] HILT vocabulary resources [EB/OL]. [2011-01-28]. <http://hilt.cdlr.strath.ac.uk/Sources/vocabulary.html>.
- [228] Hillmann D, Sutton S, Phipps J, et al. A metadata registry from vocabularies up:The NSDL registry project[C/OL]/Baker, T.& Solorio, J. Proceedings of 2006 International Conference on Dublin Core and Metadata Applications: metadata for knowledge and learning. Colima, Mexico: Dublin Core Metadata Initiative.2006:65-75.[2011-01-28]. <http://arxiv.org/ftp/cs/papers/0605/0605111.pdf>.
- [229] <http://dublincore.org/documents/2003/04/02/dc-xml-guidelines/#note1> (2008-12-29)
- [230] <http://dublincore.org/documents/usageguide/qualifiers.shtml> (2008-12-24)
- [231] <http://dublincore.org/schemas/xmls/> (2008-12-29)
- [232] <http://library.gxcedu.com/wt3.html> (2008-11-16)
- [233] http://metadata.teldap.tw/standard/standard-big5/cdwa_2-0_draft.pdf (2009-1-6)
- [234] <http://www.bl.uk/bibliographic/natbib.html>
- [235] http://www.fgdc.gov/metadata/documents/workbook_0501_bmk.pdf(2009-1-8)
- [236] http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf (2009-1-8)
- [237] http://www.getty.edu/research/conducting_research/standards/cdwa/ (2009-1-8)
- [238] <http://www.loc.gov/standards/marcxml/Sandburg/sandburg.mrc> (2008-11-17)
- [239] <http://www.loc.gov/standards/mods/mods-outline.html>(2008-12-29)
- [240] http://www.loc.gov/standards/mods/v3/mods-userguide-examples.html#journal_article (2008-12-29)
- [241] <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/> (2009-1-11)
- [242] IFLA Blogs[EB/OL]. [2013-08-22]. <http://blogs.ifla.org/>.
- [243] Isaac, A., Summers, E. SKOS Simple Knowledge Organization System Primer (W3C Working Group Draft 21 February 2008) [EB/OL]. [2011-04-10]. <http://www.w3.org/TR/skos-primer/>.
- [244] Isaac, Antoine, Ed Summers.SKOS Simple Knowledge Organization System Primer [EB/OL]. [2008-06-20]. <http://www.w3.org/TR/skos-primer/>.
- [245] IFLA UNIMARC Core Activity Publications. <http://www.ifla.org/VI/8/unimarc-publist.htm> (2008-11-16)
- [246] Jan é e G, Ikeda S, Hill L. The ADL thesaurus protocol (version 1.0)[OL].[2011-01-28]. <http://www.alexandria.ucsb.edu/thesaurus/specification.html>.
- [247] Leader(NR).<http://www.loc.gov/marc/bibliographic/bdleader.html>(2008-11-17)
- [248] Lexurus bank[OL].[2011-01-28].<http://www.vocman.com/lexaurusbank>.
- [249] MARC 21 Specification for Record Structure, Character Set, and Exchange Media. <http://www.loc.gov/marc/specifications/specrecstruc.html> (2008-12-2)
- [250] MARC in XML. <http://www.loc.gov/marc/marcxml.html#marcdtd>(2008-12-5)
- [251] MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media-RECORD STRUCTURE. <http://www.loc.gov/marc/specifications/specrecstruc.html>(2008-11-17)
- [252] MARC XML Architecture. <http://www.loc.gov/standards/marcxml/marcxml-architecture.html>(2008-11-17)
- [253] Metadata. <http://en.wikipedia.org/wiki/Metadata>(2008-12-7)
- [254] Mike Pullin. Welcome to a Z39.50 Instructional Site[OL/EB]. [2008-11-10]. <http://www.txmike.com/Presentations/Z3950/>
- [255] OIL[EB/OL].[2013-06-10]. <http://www.ontoknowledge.org/oil>.
- [256] OWL Web Ontology Language Semantics and Abstract Syntax[EB/OL].(2009-11-12) [2013-06-10].

- <http://www.w3.org/TR/owl-semantic>.
- [257] OWL[EB/OL].[2013-06-10].<http://www.w3.org/owl>.
- [258] 00X - Control Fields-General Information.<http://www.loc.gov/marc/bibliographic/bd00x.html> (2008-12-2)
- [259] Platform for Internet Content Selection (PICS). [http://www.w3.org/PICS/#Specs\(2009-01-07\)](http://www.w3.org/PICS/#Specs(2009-01-07))
- [260] Pubby.[EB/OL].[2010-07-01].<http://wifo5-03.informatik.uni-mannheim.de/pubby/>
- [261] RDF Schema 1.1[EB/OL].[2013-06-10].<http://www.w3.org/TR/rdf-schema/>.
- [262] RDF[EB/OL]. [2013-06-30]. <http://www.w3.org/RDF/>.
- [263] RDF 和 XML 的区别[EB/OL].[2013-06-01]. http://www.360doc.com/content/10/0426/10/865714_24925727.shtml
- [264] Sauermann, Leo, Richard Cyganiak. Cool UR Is for the Semantic Web [EB/OL]. [2008-06-20]. <http://www.w3.org/TR/cooluris/>.
- [265] SKOS Simple Knowledge Organization System Reference[EB/OL].(2008-01-25)[2013-06-10]. <http://www.w3.org/TR/2008/WD-skos-reference-20080125/>
- [266] SKOS Simple Knowledge Organization System Reference[EB/OL].(2009-08-18)[2013-06-10]. <http://www.w3.org/TR/skos-reference/#xl>
- [267] SKOS Simple Knowledge Organization System Reference[EB/OL].(2009-08-18)[2013-06-10]. <http://www.w3.org/TR/skos-reference/#xl>
- [268] SKOS Simple Knowledge Organization System-Home Page[EB/OL].[2013-06-10]. <http://www.w3.org/2004/02/skos/>.
- [269] Stephenson M.Indexing resources on the WWW:database indexing, controlled vocabularies & thesauri[EB/OL][2011-01-28].<http://www.slais.ubc.ca/resources/indexing/database1.htm>.
- [270] Taxonomy warehouse[OL].[2011-01-28].<http://www.taxonomywarehouse.com/>.
- [271] Tudhope D, Koch T, Heery R.Terminology services and technology:JISC state of the art review[R/OL]. UK: Joint Information Systems Committee(JISC), 2006.[2011-01-28]. http://www.jisc.ac.uk/Terminology_Services_and_Technology_Review_Sep_06.
- [272] The DARPA Agent Markup Language Homepage[EB/OL].[2013-06-10]. <http://www.daml.org>.
- [273] Using Dublin Core. <http://dublincore.org/documents/usageguide/index.shtml>(2008-12-25)
- [274] Vizine-Goetz D. Terminology services[EB/OL].[2011-01-08]. <http://www.oclc.org/research/presentations/vizine-goetz/cendi-nkos-isko.ppt>.
- [275] Web of KnowledgeSM[EB/OL]. [2013-08-13]. <http://wokinfo.com/>.
- [276] Web of Science 索引数据库介绍, available from: <http://www.isiproducts.com>
- [277] Wilson 公司索引数据库介绍, available from: <http://www.hwwilson.com>
- [278] XML 认证教程, 第6部分: XML Schema. <http://www.ibm.com/developerworks/cn/xml/x-cert/part6/index.html>(2008-12-7)